Language Detection and Tracking in Multilingual Documents Using Weak Estimators[*]

Aleksander Stensby[1,**], B. John Oommen[1,2], and Ole-Christoffer Granmo[1]

[1] Dept. of ICT, University of Agder, Grimstad, Norway
[2] School of Computer Science, Carleton University, Ottawa, Canada[***]

Abstract. This paper deals with the extremely complicated problem of language detection and tracking in real-life electronic (for example, in Word-of-Mouth (WoM)) applications, where various segments of the text are written in different languages. The difficulties in solving the problem are many-fold. First of all, the analyst has no knowledge of when one language stops and when the next starts. Further, the features which one uses for any one language (for example, the n-grams) will not be valid to recognize another. Finally, and most importantly, in most real-life applications, such as in WoM, the fragments of text available before the switching, are so *small* that it renders any meaningful classification using traditional estimation methods almost meaningless. Earlier, the authors of [10] had recommended that for a variety of problems, the use of strong estimators (i.e., estimators that converge with probability 1) is sub-optimal. In this vein, we propose to solve the current problem using novel estimators that are pertinent for non-stationary environments. The classification results which involve as many as 8 languages demonstrates that our proposed methodology is both powerful and efficient.

Keywords: Multilingual language detection, Weak estimators.

1 Introduction

In this paper, we consider the fascinating problem of language detection and tracking in real-life electronic (for example, in Word-of-Mouth (WoM)) applications. Unlike more traditional Pattern Recognition (PR) problems, in this case we encounter the scenario where the various segments of the text are written in different languages, and are both short and "chatty". We know that every PR

[*] The second author was partially supported by NSERC, the Natural Sciences and Engineering Research Council of Canada. All the authors gratefully acknowledge the support of *Integrasco*, a Word-of-Mouth company whose headquarters is in Grimstad, Norway, and in particular of its CEO, Mr. Jan Hansen.

[**] Former student. Can be contacted at: C/o Dr. Ole-Christoffer Granmo, Dept. of ICT, University of Agder, Grooseveien 36, 4876 Grimstad, Norway.

[***] *Chancellor's Professor*; *Fellow : IEEE* and *Fellow : IAPR*. The Author also holds an *Adjunct Professorship* with the Dept. of ICT, University of Agder, Norway.

E.R. Hancock et al. (Eds.): SSPR & SPR 2010, LNCS 6218, pp. 600–609, 2010.

problem essentially involves two issues, namely the training and the classification of the patterns. In the training phase, the class-conditional distribution of the features is estimated, based on the given training samples. Generally speaking, traditional PR systems assume that the class-conditional distributions are stationary, and thus that they do not change with time. However, in the case of the problem we study, as we shall see, the training data possesses non-stationary class-conditional distributions. All of these issues render the problem being studied both difficult and non-trivial.

The traditional strategy to deal with non-stationary environments has been one of using a *sliding window* [6]. The problem with this is that if the size of the window is too small, the corresponding estimates tend to be poor. If one chooses a too-large window size, the estimates prior to the change of the parameter have too much influence on the new estimates. Also, the observations during the entire window width must be maintained and updated during the process of estimation.

There are numerous problems which have been recently reported, where strong estimators pose a real-life concern. Recently Oommen and his co-authors presented a strategy by which the parameters of a binomial/multinomial distribution can be estimated when the distribution is non-stationary [10]. The method is referred to as the *Stochastic Learning Weak Estimator* (SLWE), and is a novel estimation method based on the principles of stochastic learning. We propose to use the SLWE in our particular PR problem.

1.1 Topic Detection and Tracking and Word of Mouth

The non-stationary phenomenon described above occurs in the PR problems related to Topic Detection and Tracking (TDT) in online discussions, where the content of the discussions represents the opinions of users from all over the world. This kind of information has high value for market-oriented or consumer-focused companies.

The phenomenon of consumers providing information to other consumers is often referred to as Word of Mouth (WoM). It turns out that the nature of these discussions, consisting of multiple opinions, different topics, and a variety of languages, presents us with a problem of designing training and classification strategies when the class-conditional distributions are non-stationary.

The main difference between classification of news articles or journal papers and WoM discussions, is that these discussions generally contain the opinions of *several* different authors. Considering a discussion where several authors write parts of it means that we have a document with continuous content changes.

Treating the whole discussion as one contiguous document, the task at hand is thus to segment the discussion and to classify each segment according to the pre-defined classes, whether it be topics, sentiment or language.

Another important aspect of text classification of such WoM discussions is that the postings often are composed on the fly by the different users, without any form of spell checking. Thus, when performing text classification on such data, one must tolerate the presence of different kinds of textual errors, such as spelling and grammatical errors. Abbreviations and Internet “slang” may also be present. The classification process must work reliably on all input, and must tolerate these kind of errors to some extent. *The complexity of the problem being studied should thus be obvious to the reader*!

1.2 Contributions of This Paper

The present work develops an efficient and accurate methodology for the training and testing of topic detection and tracking in multilingual online discussions. In contrast to the state-of-the-art, we introduce a novel approach to language classification in multilingual documents where the classification is done without any prior segmentation of the sample document, and where we do not require the class-conditional distributions of the “features” to be stationary. The method utilizes the principles of the SLWE proposed by Oommen *et al.* to update the probabilities of the input samples, combined with mixed-order n-grams as the discriminatory features, based on an n-gram language model [4]. In the light of the above, we believe that our work is both novel and of a pioneering sort.

2 Language Classification in Mono/Multilingual Documents

A crucial problem that has received little attention in the literature is that of classifying documents containing several languages, or so-called multilingual documents. The task of language classification has been widely studied, but most of the approaches focus on classifying documents written in a single language, often referred to as monolingual documents.

There are several different approaches to selecting features for language identification. These include, for instance, the presence of particular characters as discriminators [13] or the presence of particular character n-grams [12]. Cavnar and Trenkle approached the task of language classification in monolingual documents in [1], by using n-gram analysis.

Other frequently used approaches to language classification are the dictionary approach or use of words that commonly appear in the language of interest[5]. Such non-linguistically motivated features generally perform well for documents of moderate length, but their performance is significantly decreased when the length of the sample text gets shorter. Other approaches to language classification using linguistic factors that differ among languages are also found in the literature. One such approach is based on the use of *morphological* features presented by Creutz in [3] and [2]. The problem with these approaches is that the construction of a morphological lexicon for a given language requires a large amount of work by trained experts.

With respect to multilingual documents, Ozbek *et al.* presented an approach in [11], where they make use of the Creutz algorithm. Their approach demonstrated good results for the Turkish language, but the results were discouraging for the English language, with a worst case accuracy of 40%. Ludovik and

Zacharski proposed an algorithm for classifying multilingual documents that is based on mixed-order n-grams, Markov chains, maximum likelihood and dynamic programming in [7]. Language classification in multilingual documents using a word-window approach was studied in [8] by Mandl *et al.* Their results demonstrated a high accuracy for detecting the languages, but they pointed out that determining the location of the language shift was the hardest challenge, reporting a cumulative precision of 81% for locating the change point with at most 2 words off the real change point.

Our proposed method is distinct from all of the above. We are interested in classification tasks that involve the non-stationarity found in such multilingual documents, in which moreover, we do not require the scheme to know the boundaries of the different language segments in the document.

3 Weak Estimators: The SLWE

The fundamental estimation strategy that we advocate for the problem being studied is the SLWE alluded to earlier. We shall explain it, in some detail, here.

When dealing with an alphabet of r symbols, whose probabilities have to be estimated "on the fly", the best model is to assume that the input symbol is drawn from a multinomial random variable. The multinomial distribution is characterized by two parameters, namely, the *number* of trials, and a probability vector which determines the probability of a specific event (from a pre-specified set of events) occurring. In this regard, we assume that the number of observations is the number of trials. Therefore, the problem is to estimate the latter probability *vector* associated with the set of possible outcomes or trials.

Specifically, let X be a multinomially distributed random variable, which takes on the values from the set $\{`1', \ldots, `r'\}$. We assume that X is governed by the distribution $S = [s_1, \ldots, s_r]^T$ as $X = `i'$ with probability s_i, where $\sum_{i=1}^{r} s_i = 1$. Also, let $x(n)$ be a concrete realization of X at time 'n'. The intention of the exercise is to estimate S, i.e., s_i for $i = 1, \ldots, r$. We achieve this by maintaining a running estimate $P(n) = [p_1(n), \ldots, p_r(n)]^T$ of S, where $p_i(n)$ is the estimate of s_i at time 'n', for $i = 1, \ldots, r$, with $\sum_{i=1}^{r} p_i(n) = 1$. Then, the value of $p_1(n)$ is updated as per the following simple rule (the rules for other values of $p_j(n)$ are similar):

$$p_1(n+1) \leftarrow p_1 + (1-\lambda)\sum_{j\neq 1} p_j \quad \text{when } x(n) = 1 \tag{1}$$

$$p_1(n+1) \leftarrow \lambda p_1 \quad \text{when } x(n) \neq 1 \tag{2}$$

The vector $P(n) = [p_1(n), p_2(n), \ldots, p_r(n)]^T$ refers to the estimate of $S = [s_1, s_2, \ldots, s_r]^T$ at time 'n', and we will omit the reference to time 'n' in $P(n)$ whenever there is no confusion. The above updating rules, with $\lambda \in [0, 1]$ being the learning rate, lead to asymptotic values of P whose mean converges exactly to S. The proof of this property and the properties concerning the variance and convergence of the limiting distribution are found in [9].

4 SLWE Solution to Language Detection and Tracking

By combining the SLWE with mixed-order n-gram models, we present a novel approach to the task of language classification in multilingual WoM documents.

One important issue in all PR systems is that of selecting the feature space of the classifier. The approach we advocate is akin to the ideas of Cavnar and Trenkle, which uses mixed-order n-gram models, and builds n-gram profiles for each language that is being classified. The nature of WoM discussions were also a key motivating factor in choosing n-grams as features, due to their robustness with regard to noise in the input text and that the segments may be too short for word-based features to encapsulate sufficient information.

By utilizing n-grams, there is no need for preprocessing in the sense of spell checking or stemming since n-grams essentially gives us the information-bearing content of a word without performing such costly procedures. In addition, stemming requires sophisticated knowledge about the language, and is thus useless for our task since we do not know the language of the input text. The SLWE also possesses better scalability than, for instance, the MLE, which is used by Ludovik *et al.* [7] in their approach, with regard to a large number of features. Another important motivation for using the SLWE for this task is that there is no need for a separate segmentation process by using complex methods such as dynamic programming used by Ludovik and his co-authors [7]. Instead, the SLWE is able to adapt to changes quickly if the environment switches its probability vector, which in our case is the distribution of top n-grams for the possible languages being classified.

4.1 The Basic Algorithm

The PR system presented here for classification of language in multilingual documents, consisted of two phases. The first phase involved training mixed-order n-gram profiles for each language that the system should support. Only the most frequent n-grams of order $n = 1$ to 4 for a given language were kept in the profile. The second phase of the PR system consisted of the actual classification, or testing phase. In this phase, the estimate of the SLWE was initialized at the beginning of each document, with a feature vector consisting of all unique n-grams from each of the different language profiles. Each document in the testing corpus was processed, and for each document, each word was processed and classified according to a distance measure between the estimated probability vector and each of the language probability distributions. The running estimate of the SLWE was updated after every word was processed.

Training Language Profiles. The training set consisted of monolingual documents, pre-labeled with the language they were written in. Each document in this training set was subjected to a tokenization process. We also removed all non alphanumerical characters from the text. After the tokenization process was done, each word in the document was expanded to their mixed-order n-grams.

After all the n-grams were read, the frequencies were converted into probabilities by dividing each frequency by the total number of observed n-grams. By doing so, we were able to obtain an n-gram probability distribution for the given language.

Classification and Testing. The second phase consisted of classifying each document in the testing corpus, using the SLWE and the probability distributions for each language.

The test documents were generated by our system, by concatenating segments from monolingual documents. This approach made it possible for us to pre-label each segment of the multilingual sample document, allowing us to validate the classification results for each segment.

Each document to be classified was read into the system and was subjected to the same tokenization process as described for the training phase. The feature vector of the SLWE consisted of all the unique n-grams from all the language profiles defined for the system. The SLWE kept a running estimate of this feature vector, where each n-gram was associated with a given probability. These probabilities were initialized evenly.

After the SLWE was initialized, and the document was tokenized into a list of words, the system was ready to perform the actual classification procedure. The formal algorithm is included in the unabridged paper and omitted here due to space limitations.

For each of the words that the sample document contains, the system expanded the word into mixed-order n-grams. Then, for each of these n-grams, the probabilities of the running estimate was updated as per the multinomial updating scheme of the SLWE. If the n-gram is found in the estimate probability vector, its probability was increased according to the updating rules. The probability of all other n-grams were then accordingly reduced. If the n-gram were not in the estimate vector, it was merely ignored.

After all the n-grams for the given word were processed, the system measured a distance between the estimated probability vector and each of the language probability distributions. The word was then classified as being written in the language represented by the language profile that measured the shortest distance from the estimate (using the distance measure alluded to earlier). With the assumption that a sentence is monolingual, we counted the number of words in a sentence and classified the sentence as being written in the language that had the highest word classification count. The validation results are maintained in a so-called *confusion matrix.*

5 Experimental Results

The motivation for these experiments was to investigate how well our algorithm was suited for language classification in multilingual documents, and by testing several different languages we sought to investigate the ability to classify documents written in different languages and how well the classifier would scale with regard to the number of supported languages. We use different values for the cut-off threshold to examine how well the classifier scaled with regard to the number of features, and we experimented with different values for the learning parameter of the SLWE to evaluate the impact of slow versus fast convergence when dealing with language classification. We also measured the accuracy of our classifier operating with different sentence lengths to see how well it is able to deal with short or long sentences.

5.1 Experimental Setup

The classifier was tested on three different sets of languages, generated by concatenating sentences from monolingual documents. The languages used for our testing are English, French, and German for Experiment Set 1, and English English, French, German, Norwegian, Italian, Spanish, Dutch and Swedish for Experiment Set 3. Details of Experiment Set 2 can be found in the unabridged version of this paper. For each of these sets we generated different variants using different sentence lengths. All test sets had a corpus size of 100 documents, except for test set VI which had 200 documents. Test set I, II and III, for experiment set 1, consisted of respectively 10, 15 and 20 words per sentence. The final test set, VI, for experiment set 3, contained 20 words per sentence.

With these test sets we tested our classifier on four different test cases, using different values for the learning parameter, λ, and different cut-off thresholds. Test case A and B used a cut-off threshold of 400, whereas test case C and D

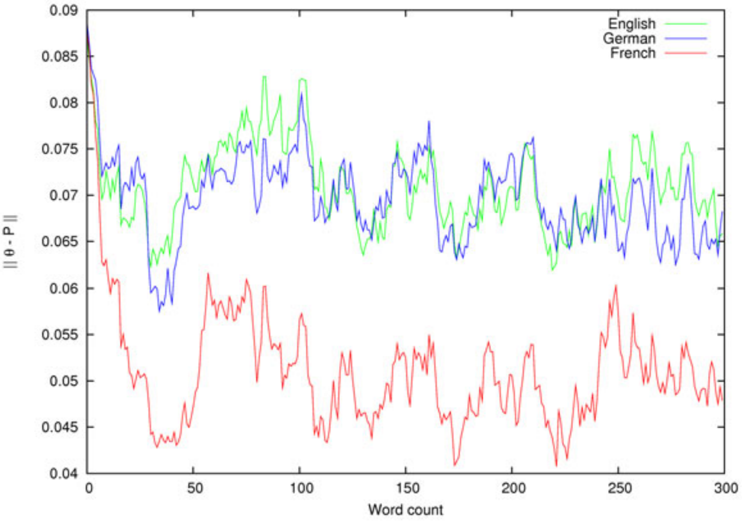

Fig. 1. Plot of the Euclidean distance from the estimated probability vector to each of the language profiles. The document being classified was monolingual, written in French, containing 300 words.

used 500 as the cut-off threshold. For the learning parameter, λ, test case A and C used a value of 0.98. Test case B and D used 0.99 as the learning parameter. Figure 1 shows a plot of the Euclidean distance between the estimate $P(n)$ and three possible language profiles for a document that is monolingual. Despite the document being monolingual, the system assumes that the document is multilingual. The sample being classified contains 300 words written in French and in this example, the classifier operates on word-level, disregarding any sentence boundaries. We observe that the SLWE converges rapidly to the true language profile, which for this sample was French. Even though the variance of the estimate is rather high, we observe that the distance to the other language profiles is far greater than the distance to the correct language profile. We used $\lambda = 0.99$ and 300 as the cut-off threshold in this experiment.

5.2 Results

Language Set 1. The classification accuracy for our first language set is reported for each of the test cases in Table 1.

Table 1. Reported classifier accuracy for each of our test cases for the first language set

Test Set	Test Case	λ	Cut-off	Accuracy (Eng)	Accuracy (Fre)	Accuracy (Ger)
I.	A	0.98	400	0.968	0.962	0.949
I.	B	0.99	400	0.941	0.891	0.920
I.	C	0.98	500	0.970	0.960	0.949
I.	D	0.99	500	0.945	0.905	0.925
II.	A	0.98	400	0.973	0.990	0.987
II.	B	0.99	400	0.951	0.963	0.966
II.	C	0.98	500	0.971	0.992	0.987
II.	D	0.99	500	0.961	0.965	0.974
III.	A	0.98	400	0.996	0.990	0.983
III.	B	0.99	400	0,987	0.986	0.974
III.	C	0.98	500	0.994	0.990	0.983
III.	D	0.99	500	0.988	0.986	0.974

We observe that best accuracy for all the test sets is achieved with the learning parameter λ set to 0.98. Higher values of λ yields slower, but more accurate convergence. When classifying short sentences, it is important that the SLWE is able to converge rather quickly so that as few words as possible in the sentences are misclassified. We also observe that the different cut-off thresholds only to a small extend affects the classifier accuracy.

Table 2 shows the confusion matrix for test case A on test set III, which demonstrated an averaged classifier accuracy of 0.9896. In this experiment, the test set consisted of 520 sentences in English, 515 sentences in French and 465 sentences in German. Each sentence consists of 20 words. By looking at the accuracies listed in Table 2, we observe that only two of the 520 sentences in English were misclassified. One of these as French and the other as German.

Table 2. Confusion matrix for test case A, using test set III

	Eng	**Fre**	**Ger**
Eng	0.996	0.002	0.002
Fre	0.010	0.990	0.000
Ger	0.013	0.004	0.983

Language Set 3. For the last language set we tested our classifier using all eight languages that we had generated language profiles for. For this case we generated the test samples using a sentence length of 20 words. This testing corpus consisted of 200 documents, and the results are listed in Table 3.

Table 3. Reported classifier accuracy for each of our test cases for the third language set with eight different languages

Test Set	**Test Case**	λ	**Cut-off**	**Averaged Acc.**	**Best Acc.**	**Worst Acc.**)
VI.	A	0.98	400	0.9695	0.988 (Fre)	0.928 (Nor)
VI.	B	0.99	400	0.9701	0.986 (Ita)	0.928 (Nor)
VI.	C	0.98	500	0.9690	0.988 (Fre)	0.916 (Nor)
VI.	D	0.99	500	0.9717	0.986 (Ita)	0.931 (Nor)

5.3 Discussion and Summary of Results

We have observed that our classifier is able to classify multilingual documents with high overall accuracy. Our experiments demonstrates that the classifier performs extremely well for moderate-sized segments, and that it performs adequately for shorter sentences with 10 words per sentence.

For the first language set, we obtained a classification accuracy for the English language as high as 0.996 using $\lambda = 0.98$ and the cut-off threshold set to 400. This accuracy was achieved with sentences consisting of 20 words. For shorter segments, with 10 words per sentence, we achieved an accuracy of 0.97. This is still a fairly good accuracy considering the length of the segments. We observe that using a cut-off threshold around 400 yields satisfying results, which is in accordance to the suggested cut-off thresholds used by Cavnar and Trenkle in their experiments. This also shows us that by reducing or increasing the feature space, the classifier scales well and is not notably handicapped by working with a limited feature set compared to a larger one.

For the last language set, using eight different languages, we observed through our experiments that our classifier is able to scale well with regard to the number of supported languages. The averaged accuracy reported for our experiments was slightly lower than for the case when dealing with only five languages, but the classifier still performs well with an error rate of only 0.0283 for eight languages, compared to an error rate of 0.0186 in the case of five languages.

6 Conclusion and Future Work

In this paper we have studied the problems of topic detection and tracking in multilingual online discussions, which is particularly difficult because the content involve the brief and “chatty” opinions of users in multiple languages. Unlike the traditional PR problem, in this scenario, the class-conditional distributions are non-stationary. By using the estimation philosophy recommended in [10], we have proposed a solution to the current problem using novel estimators that are pertinent for non-stationary environments. The classification results obtained for various data sets which involve as many as 8 languages demonstrates that our proposed methodology is both powerful and efficient.

References

1. Cavnar, W.B., Trenkle, J.M.: N-gram-based text categorization. In: Proceedings of SDAIR 1994, 3rd Annual Symposium on Document Analysis and Information Retrieval, Las Vegas, US, pp. 161–175 (1994)
2. Creutz, M.: Unsupervised segmentation of words using prior distributions of morph length and frequency. In: ACL 2003: Proceedings of the 41st Annual Meeting on Association for Computational Linguistics, Morristown, NJ, USA, pp. 280–287. Association for Computational Linguistics (2003)
3. Creutz, M., Lagus, K.: Unsupervised discovery of morphemes (2002)
4. Dunning, T.: Statistical Identification of Language. Technical report MCCS 94-273. New Mexico State University (1994)
5. Ingle, N.C.: A language identification table. The Incorporated Linguist. 15(4), 98–101 (1976)
6. Jang, Y.M.: Estimation and Prediction-Based Connection Admission Control in Broadband Satellite Systems. ETRI Journal 22(4), 40–50 (2000)
7. Ludovik, Y., Zacharski, R.: Multilingual document language recognition for creating corpora. Technical report, New Mexico State University (1999)
8. Mandl, T., Shramko, M., Tartakovski, O., Womser-Hacker, C.: Language identification in multi-lingual web-documents. In: Kop, C., Fliedl, G., Mayr, H.C., Métais, E. (eds.) NLDB 2006. LNCS, vol. 3999, pp. 153–163. Springer, Heidelberg (2006)
9. Oommen, B.J., Rueda, L.: Stochastic Learning-based Weak Estimation of Multinomial Random Variables and Its Applications to Non-stationary Environments. Pattern Recognition (2006) (in Press)
10. Oommen, B.J., Rueda, L.: Stochastic Learning-based Weak Estimation of Multinomial Random Variables and Its Applications to Non-stationary Environments. Pattern Recognition 39(1), 328–341 (2006)
11. Ozbek, G., Rosenn, I., Yeh, E.: Language classification in multilingual documents. Technical report, Stanford University (2006)
12. Souter, C., Churcher, G., Hayes, J., Hughes, J., Johnson, S.: Natural language identification using corpus-based models. Hermes Journal of Linguistics 13, 183–203 (1994)
13. Ziegler, D.: The automatic identification of languages using linguistic recognition signals. PhD thesis, Buffalo, NY, USA (1991)