Tweets language identification using feature weighting

Identificación de idioma en tweets mediante pesado de términos

Juglar Díaz Zamora
Universidad de Oriente
Santiago de Cuba, Cuba
juglar.diaz@cerpamid.co.cu

Adrian Fonseca Bruzón
Reynier Ortega Bueno
CERPAMID
Santiago de Cuba, Cuba
{adrian, reynier.ortega}@cerpamid.co.cu

Resumen: Este trabajo describe un método de detección de idiomas presentado en el Taller de Identificación de Idioma en Twitter (TweetLID-2014). El método propuesto representa los tweets por medio de trigramas de caracteres pesados de acuerdo a su relevancia para cada idioma. Para el pesado de los trigramas se emplearon tres esquemas de pesados de características tradicionalmente usados para la reducción de la dimensionalidad en la Clasificación de Textos. El idioma de cada tweet se obtiene mediante mayoría simple luego de sumar los pesos que cada trigrama presente en el tweet aporta para cada idioma. Finalmente, realizamos un análisis de los resultados obtenidos.
Palabras clave: tweets, identificación de idioma, pesado de rasgos

Abstract: This paper describes the language identification method presented in Twitter Language Identification Workshop (TweetLID-2014). The proposed method represents tweets by weighted character-level trigrams. We employed three different weighting schemes used in Text Categorization to obtain a numerical value that represents the relation between trigrams and languages. For each language, we add up the importance of each trigram. Afterward, tweet language is determined by simple majority voting. Finally, we analyze the results.
Keywords: tweets, language identification, feature weighting

1 Introduction

With the growing interest in social networks like Twitter and Facebook, the research community has focused on applying data mining techniques to such sources of information. One of these sources are the messages produced in the social network Twitter, known as tweets. Tweets represent a challenge to traditional Text Mining techniques mainly due to two characteristics, the length of the texts (only 140 characters allowed) and the Internet Slang present in these texts. Because of the limitations of 140 characters, people create their own informal linguistic style by shortening words and using acronyms.

Language identification (LID) is the task of identifying the language in which a text is written. This is an important pre-processing step necessary for traditional Text Mining techniques; also, Natural Language Processing tasks like machine translation, part of speech tagging and parsing are language dependent.

Some work has been done for LID in long and well-formed texts; traditional approaches are focused on words and n-grams (of characters). In word-based features, we can find short word-based and frequency-based approaches. (Grefenstette, 1995) proposed a short word-based approach where he uses words up to five characters that occurred at least three times. The idea behind this approach is the language specific significance of common words like conjunctions having mostly only marginal lengths. In frequency based approach, (Souter et al., 1994) takes into account one hundred high frequent words per language extracted from training data for 9 languages and 91% of all documents were correctly identified.

The n-gram based approach uses n-grams of different (Cavnar and Trenkle, 1994) or fixed (Grefenstette, 1995; Prager, 1999) lengths from tokenized words.

(Cavnar and Trenkle, 1994) evaluate their algorithm on a corpus of 3713 documents in 14 languages, for language models of more than

300 n-grams very good results of 99.8% were achieved.

The n-gram technique described by (Grefenstette, 1995) calculates the frequency of trigrams in a language sample, the probability of a trigram for a language is approximated by summing the frequency of all trigrams for the language and dividing the trigram frequency by the sum of all frequencies in the language. The probabilities are then used to guess the language by dividing the test into trigrams and calculating the probability of the sequence of trigrams for each language, assigning a minimal probability to trigrams without assigned probabilities. The language with the highest probability for the sequence of trigrams is chosen.

We consider that for LID to be effective, the inflected forms of a root word should being related to the same word, and knowing that the character-level n-grams of different morphological variations of a word tend to produce many of the same n-grams, we chose to use character-level n-grams as features in our approach. Since trigrams have proven good results in LID (Grefenstette, 1995; Prager, 1999), this is our n-grams selection.

Some studies have shown that system designed for other types of texts perform well on tweet language identification (TLID) (Lui and Baldwin, 2012), but some systems which were specifically designed for the characteristics of tweets performed better.(Carter et al., 2013).

There is also a body of work in TLID employing different techniques, for example graph representation of languages based in trigrams (Trompand and Pechenizkiy, 2011), combination of systems (Carter et al., 2013), user language profile, links and hashtags (Carter et al., 2013).

We propose a language identification system based on feature weighting schemes (FWS), commonly used in Text Categorization (TC). We obtain a numerical value that represents the relation between features, trigrams of characters in our case, and languages. This proposal can be extended to words and longer or shorter n-grams.

The remainder of the paper is structured as follows. In Section 2, we describe our tweet language identification system (Cerpamid-TLID2014) and the feature weighting schemes tested. In Section 3, we present experiments conducted for estimating the parameters of our system and we analyze the effect of feature weighting schemes in tweets language identification. Finally, conclusions and attractive directions for future work are exposed.

2 System description

In this section, we describe our system and the feature weighting schemes that we used in our experiments.

2.1 Feature weighting

Dimensionality reduction (DR) is an important step in Text Categorization. It can be defined as the task of reducing the dimensionality of the traditional vector space representation for documents; these are two main approaches to this task (Sebastiani, 2002):

- Dimensionality reduction by feature selection (John, Kohavi and Pfleger, 1999): the chosen features *r'* are a subset of the original *r* features (e.g. words, phrases, stems, lemmas).
- Dimensionality reduction by feature extraction: chosen features are not a subset of the original *r* features, but are obtained by combinations or transformations of the original ones.

There are two distinct ways of viewing DR, depending on whether the task is performed locally (i.e., for each individual category) or globally.

We focus on local feature selection schemes, since our interest is to obtain the importance of every trigram (features) for every language (categories).

Many locally feature selection techniques have been tried. We show in Table 1 those used in this paper, GSS Coefficient (GSS) (Galavotti, Sebastiani, and Simi, 2000), NGL Coefficient (NGL) (Ng, Goh, and Low, 1997) and Mutual Information (MI) (Battiti, 1994).

FWS	Mathematical form
MI	$\log \frac{P(t_k, c_i)}{P(t_k) * P(c_i)}$
NGL	$\frac{\sqrt{N} * [P(t_k, c_i) * P(\overline{t_k}, \overline{c_i}) - P(t_k, \overline{c_i}) * P(\overline{t_k}, c_i)}{\sqrt{P(t_k) * P(c_i) * P(\overline{t_k}) * P(\overline{c_i})}}$
GSS	$P(t_k, c_i) * P(\overline{t_k}, \overline{c_i}) - P(t_k, \overline{c_i}) * P(\overline{t_k}, c_i)$

Table 1: Feature weighting schemes.

In our case, in order to make the feature weighting schemes depending of the available amount of text of each language, and not of the number of documents. The probabilities in Table 1 are interpreted on an event space of features (Sebastiani, 2002) (e.g., $P(\overline{t_k}, c_i)$ denotes the join probability that the trigram t_k does not occur in the language c_i, computed as rate between the number of trigram in c_i different to t_k and the total number of trigrams in corpus N).

For each language, we keep the most important trigrams and discard the rest.

2.2 Language identification

Our system is a three-step procedure; first, trigrams are extracted from the tweet, then a filtering phase takes place, in this phase those tweets that do not belong to the set of languages that our system identify are labeled as *other*. Finally, a language is assigned for the tweet. We present these steps in Algorithm 1.

Algorithm 1. Cerpamid-TLID2014

Consider the tweet to identify the language, c the content of t and L_j the list of weighted trigrams for language j.

Step 1: Split *c* in trigrams

a) Split c in words.
b) Remove numbers, punctuation marks and make all the text lowercase.
c) Add underscore in the beginning and the ending of every word.
d) Obtain the list *lt* of trigrams that represent *t*.

Step 2: Filtering

a) Let *trigrams_c* be the number of trigrams in *lt* and *trigrams_in* the number of trigrams in *lt* that appear in any L_j.
b) Let $n = \frac{trigrams\_in}{trigrams\_c}$ and θ a threshold of known trigrams in c.
c) If $n > \theta$ go to step 3, else set language as *other*.

Step 3: Selecting Language

a) For each language L_j:
   i) $vote(L_j) = \sum_{t_i \in lt} weight(t_i, L_j)$
b) Label t with the most voted language.

3 Experiments

In this section we explain the estimation of threshold θ *(see section 2.1, Algorithm 1, Step 2)*, the experiments over the feature weighting schemes and results of our proposal in TweetLID-2014. In order to evaluate the feature weighting schemes and to estimate the threshold of filtering θ, the corpus provided by the organizers of TweetLID-2014 was divided into training, test and development sets. The training set is the 70% of the tweets not labeled as language *undefined* or *other*. The remaining 30% was divided again in 70% and 30%.This 30% is our development set, this last 70% and the tweets labeled *undefined* or *other* form the test set.

In TweetLID-2014 the organizers proposed two modalities; constrained, where the training only can be performed over the training set provided at TweetLID-2014 and free training (unconstrained) where is possible to use any other resource. For the free training mode, we decided increase the amount of text per language provided at TweetLID-2014 in order to provide our proposal with greater ability to differentiate one language from another. For English (161 mb), Portuguese (174mb) and Spanish (174 mb) we used texts from the Europarl corpus (Koehn, 2005), and for Catalan (650 mb), Basque (181 mb) and Galician (157 mb), articles from Wikipedia. The training corpus for our experiments in the free training mode is the 70% of the tweets not labeled as language *undefined* or *other* added to the documents from Europarl and Wikipedia.

3.1 Estimation of threshold θ

For estimating θ, first we obtain the list of trigrams weighted from the training set, even when the weights are not used in this stage. Later we obtain a list L of the values n *(see section 2.1, Algorithm 1, Step 2)* for all the tweets in the development set. Then, we repeat 10000 times a sampling with replacement over L, in every one of these iterations we select the lowest value different from zero. These values are averaged and that is our threshold θ. The idea is to estimate statistically the value of n for a tweet written in one of the languages that we identify. The value obtained in our experiment was 0.9, and this was used for all runs.

3.2 Selecting the best feature weighting scheme

In Table 2 we show the results obtained with each feature weighting scheme in our test sets, in the two modalities, constrained and unconstrained. The best FWS in both modes was MI, while for every FWS the constrained version obtained better results for LID.

In order to make a deeper analysis of our system, we show in Table 3 the precision and the numbers of assignations to every language for our best combinations of feature weighting scheme and task mode (MI in Constrained Mode).

FWS	Precision (Averaged)	Mode
MI	0.704	Constrained
MI	0.632	Free Training
NGL	0.691	Constrained
NGL	0.522	Free Training
GSS	0.585	Constrained
GSS	0.431	Free Training

Table 2: Results of each feature weighting scheme.

Language	#Tweets	Precision
English	218	0.784
Portuguese	548	0.833
Catalan	345	0.817
Other	43	0.418
Basque	109	0.623
Galician	117	0.572
Spanish	1826	0.882

Table 3: Results by language using MI in constrained mode.

3.3 Results at TLID-2014

For our participation at TLID-2014 we used the full corpus provided by the organizers and, in addition, the documents extracted from Europarl and Wikipedia for the free training mode. In Table 4 and Table 5 we show our results at TweetLID-2014. As can be seen we placed 8th between 12 about runs and 5th between 7 about groups in the constrained mode (Table 4). Our results in precision at TweetLID-2014 are similar to the results in precision that we obtained with our own test set, while the F1 measure was dropped for the lows values in recall. About the unconstrained version, we placed last with our two runs; almost all team did worst in this mode.

Group	P	R	F1
UPV (2)	0.825	0.744	0.752
UPV (1)	0.824	0.730	0.745
UB / UPC	0.777	0.719	0.736
Citius (1)	0.824	0.685	0.726
RAE (2)	0.813	0.648	0.711
RAE (1)	0.818	0.645	0.710
Citius (2)	0.689	0.772	0.710
CERPAMID (1)	0.716	0.681	0.666
UDC / LYS (1)	0.732	0.734	0.638
IIT-BHU	0.605	0.670	0.615
CERPAMID (2)	0.704	0.578	0.605
UDC / LYS (2)	0.610	0.582	0.498

Table 4: Results at TLID-2014 for constrained mode

Group	P	R	F1
Citius (1)	0.802	0.748	0.753
UPV (2)	0.737	0.723	0.697
UPV (1)	0.742	0.686	0.684
Citius(2)	0.696	0.659	0.655
UDC / LYS (1)	0.682	0.688	0.581
UB / UPC	0.598	0.625	0.578
UDC/ LYS (2)	0.588	0.590	0.571
CERPAMID (1)	0.694	0.461	0.506
CERPAMID (2)	0.583	0.537	0.501

Table 5: Results at TLID-2014 for free training mode.

4 Conclusions and future work

We presented a tweet language identification system based on trigrams of characters and feature weighting schemes used for Text Categorization. One of our run placed 8th between 12 in the constrained version at TLID-2014 whilst in the free training version we placed last. Most of the system performed better in the constrained version. We found as the main weakness of our proposal the identification of tweets labeled *other*. As future work; we consider exploring other features, test others feature weighting schemes and tackle the problem of the identification of tweets labeled as *other* with the inclusion of lists of common terms used in tweets in the step of filtering.

Bibliography

Battiti, R.1994. Using mutual information for selecting features in supervised neural net learning. Neural Networks, 5(4):537–550.

Carter, S., W. Weerkamp, and MTsagkias.2013. Microblog language identification: Overcoming the limitations of short, unedited and idiomatic text. Language Resources and Evaluation, pages 1–21.

Cavnar, W., J. Trenkle. 1994. N-gram-based text categorization. In Proceedings of the Third Symposium on Document Analysis and Information Retrieval, Las Vegas, USA.

Galavotti, L., F. Sebastiani, and M. Simi. 2000. Experiments on the use of feature selection and negative evidence in automated text categorization. In Proceedings of ECDL-00, 4th European Conference on Research and Advanced Technology for Digital Libraries (Lisbon, Portugal, 2000), 59–68.

Grefenstette, G.1995. Comparing two language identification schemes. In 3rd International conference On Statistical Analysis of Textual Data.

John, G. H., R. Kohavi, and K. Pfleger, 1994. Irrelevant features and the subset selection problem. In Proceedings of ICML-94, 11th International Conference on Machine Learning (New Brunswick, NJ, 1994), 121–129.

Koehn, P., 2005. Europarl: A Parallel Corpus for Statistical Machine Translation. In: MT Summit.

Ng, H.T., W.B. Goh, and K. L. Low. 1997. Feature selection, perceptron learning, and a usability case study for text categorization. In Proceedings of SIGIR-97, 20th ACM International Conference on Research and Development in Information Retrieval (Philadelphia, PA, 1997), 67–73.

Prager, J. 1999. Linguini: Language identification for multilingual documents. In Proceedings of the 32nd Hawaii International Conference on System Sciences.

Sebastiani, F. 2002. Machine learning in automated text categorization. ACM computing surveys (CSUR), 34(1), 1-47.

Souter, G. Churcher, J. Hayes, J. Hughes, and S. Johnson. 1994. Natural language identification using corpus-based models. Hermes Journal of Linguistics, 13:183–203.

Tromp, E. and M. Pechenizkiy. 2011. Graph-based n-gram language identification on short texts. In Proceedings of Benelearn 2011, pages 27–35, The Hague, Netherlands.

Vatanen, T., J. Vayrynen., and S. Virpioja.2010. Language identification of short text segments with n-gram models. In LREC 2010, pages 3423–3430.