Towards Multilingual Automated Classification Systems

Aibek Musaev$^1$ and Calton Pu$^2$

$^1$Department of Computer Science, University of Alabama
$^2$School of Computer Science, Georgia Institute of Technology

Abstract—In this paper we propose and evaluate three approaches for automated classification of texts in over 60 languages without the need for a manually annotated dataset in those languages. All approaches are based on the randomized Explicit Semantic Analysis method using multilingual Wikipedia articles as their knowledge repository. We evaluate the proposed approaches by classifying a Twitter dataset in English and Portuguese into relevant and irrelevant items with respect to landslide as a natural disaster, where the highest achieved F1-score is 0.93. These approaches can be used in various applications where multilingual classification is needed, including multilingual disaster reporting using Social Media to improve coverage and increase confidence. As illustration, we present a demonstration that combines data from physical sensors and social networks to detect landslide events reported in English and Portuguese.

I. INTRODUCTION

In a supervised learning setting, human labels are necessary, but they may be costly to obtain [1]. This may be especially problematic in situations that attract multilingual population, for example during disasters. In this paper, we study a specific example of such situations, namely disaster reporting based on Social Media as the use of Social Media rises during disasters [2]. Note however, that the data from Social Media often contains noise due to ambiguity of the search keywords used to collect them [3].

The problem domain that we study is the classification of data collected from Twitter into items that are relevant to landslide as a natural disaster and items that are irrelevant. The following is a list of frequent examples of tweets in English that describe landslide events that involve the movement of soil:

- Heavy rains caused a landslide that destroyed houses in Taiz. 15 people from 7 families died.
- JUST IN: Mesa County increases alert level at massive Collbran landslide site: http://dpo.st/24bgWSm

However, one of the challenges involving the use of key-words to collect data from Social Media is that they are often polysemous words, which may have one or more irrelevant meanings. The following is a list of frequent examples of irrelevant topics involving the use of the word landslide:

- Elections are a numbers game. 2016 will be the biggest landslide since Reagan beat Carter. Make America Great Again.
- Is this the real life? Is this just fantasy? Caught in a landslide, No escape from reality...

Similarly, the keywords used to collect data from Social Media in Portuguese on landslide events, are also polysemous words as shown in Table I.

In this paper we propose three approaches for automated classification of texts, namely multilingual concepts, offline translation and online translation. For evaluation purposes we use a Twitter dataset collected using landslide keywords in English and Portuguese. Finally, we present a demonstration that displays the landslide data from multiple sources on a web map shown in Figure 1.

II. FRAMEWORK OVERVIEW

In this section, we present three approaches, namely Multilingual Concepts, Offline Translation and Online Translation for an automated classification of texts in virtually any language. Each approach is based on a randomized ESA classification system, which reduces Explicit Semantic Analysis (ESA) method to speed up text classification. We provide an overview of the ESA method followed by a description of the randomized ESA classification system. Finally, we describe the proposed approaches, which further extend the ESA method to support automated classification of texts in any language supported by Wikipedia.

A. ESA Overview

Explicit Semantic Analysis (ESA) is a powerful method for representing the meaning of texts in a high-dimensional space of Wikipedia articles [4]. Given a text fragment, for example, “Bernanke takes charge”, ESA generates the following top articles that are highly relevant to the input — Ben Bernanke, Federal Reserve, Chairman of the Federal Reserve, Alan Greenspan (Bernanke’s predecessor), Monetarism (an economic theory of money supply and central banking), inflation and deflation [5].

ESA splits the text into words and each word is converted to a high-dimensional real-valued vector space. In this vector space each dimension corresponds to an article in Wikipedia. The value for each dimension in the vector is computed as the strength of association between a word and each Wikipedia article. One way to define such association strength is to use a common TF-IDF approach. Finally, the centroid of the vectors representing the individual words is computed.
B. Randomized ESA

The large size of Wikipedia repository makes it computationally infeasible to apply ESA method for practical purposes due to its high dimensionality. That is why several studies have been conducted to understand or enhance ESA performance [6], [7], [8], and [9]. A random sample of Wikipedia repository as features for rapid text classification is proposed in [10]. The reduced size of Wikipedia repository is computed based on the sample size for a proportion when sampling without replacement. Given Z-score=1.96 for 95% confidence level, $N=5,322,499$, and $\varepsilon=0.02$, the sample size $n$ should be 2,400, where $N$ is the number of articles in the English Wikipedia$^1$.

C. Proposed approaches

We propose three approaches for automated classification of texts in any language, namely Multilingual Concepts, Offline Translation and Online Translation — see an overview in Table II.

### Multilingual Concepts

The ESA method represents the meaning of texts based on their association strength to Wikipedia articles. The Wikipedia articles have titles, which are the concepts described in the articles. These articles also frequently contain multilingual versions, for example the article on landslides has more than 60 versions in other languages$^2$. In other words, the same concept is represented in multiple languages. We use this fact to build ESA matrices in different languages. Note that these multilingual versions describe the same concept. Hence, our hypothesis is that a vector generated using the ESA method for a text discussing a particular concept in one language will have a similar pattern as a vector generated for a text discussing the same concept in another language. This means that given an ESA matrix based on Wikipedia articles in one language (e.g., English), we

---


$^2$https://en.wikipedia.org/wiki/Landslide
can generate a corresponding ESA matrix in another language (e.g., Portuguese) using the Portuguese version of the same articles, and apply the same classifier model that was built using the English based ESA to classify data in Portuguese. There are two advantages of this approach: there is no need for the manual annotation of a training dataset in another language, and there is no need for the translation of a training or evaluation datasets.

As shown in Table II, in the Multilingual Concepts approach, we build the randomized ESA in the original language, e.g., English, then we use it to generate vectors in the same language, so we can build a classifier model. Next we build the randomized ESA in the new language, e.g., Portuguese, apply it to generate vectors in Portuguese, and then we reuse the classifier model generated using English for classification of vectors generated using Portuguese.

**Offline Translation.** In this approach, we propose to translate the training dataset in the original language, e.g., English, to the new language, such as Portuguese. Then we build the randomized ESA in the new language, and use it to generate training vectors in the same language, so that we can build a classifier model. Finally, we use the built model for classification of vectors generated using the new language. Unlike the Multilingual Concepts approach, here we need to translate the training dataset from the original language to the new language. This is a one-time operation, which can be performed offline, so it should not affect the runtime performance of this approach. Again, there is no need for the manual annotation of the training dataset in the new language.

**Online Translation.** In this approach, we propose to translate the evaluation dataset from the new language, e.g., Portuguese, to the original language, e.g., English. Then we can reuse the classifier model that was already built for the original language. The advantage of this approach is that we do not need to annotate the training dataset in the new language. However, each time we get a new text for classification purposes, such as an incoming tweet, we need to translate it to the original language. This can become expensive both in performance and money especially given a high rate of incoming tweets.

### III. Experimental Evaluation

A. Dataset description

For experimental evaluation, we consider the performance of the proposed approaches in a specific problem domain, namely the disambiguation of landslide data collected from Social Media.

We use keywords landslide and mudslide to collect the data from Twitter in English and keywords deslizamento and desabamento to collect the data in Portuguese. Each tweet is manually labeled as either relevant or irrelevant to landslide as a natural disaster. To mark items as relevant we use two approaches. First we check whether a given text describes a landslide event confirmed by the authoritative source, namely USGS. Otherwise, we search for a confirmation of the landslide event online by other trustworthy sources using the described event as a search query.

Irrelevant items are much easier to label as we only need to check whether a given item uses other meanings of the word landslide, including an overwhelming victory in election or an excerpt from the lyrics. See an overview of the dataset used in the experiments in Table III. We use 2,400 multilingual articles in Wikipedia, which have both English and Portuguese versions. In other words, they describe the same concepts, but in two different languages. Our training dataset consists of English tweets and the evaluation dataset consists of Portuguese tweets only.

B. Experimental setup

In order to generate ESA matrices, we download 2,400 random articles from Wikipedia, such that each article has at least 100 words and has both English and Portuguese versions. We use Microsoft Translator API to translate the training dataset from English to Portuguese and the evaluation dataset from Portuguese to English. For evaluation of classification algorithms, we use the Weka software package [11]. We select SMO, Logistic Regression, Naive Bayes, J48, and Random Forest classifiers for our analysis. They represent different categories of classification algorithms and we determine the algorithms that demonstrate the best performance for each of the proposed approaches.

C. Results

The performance of the Multilingual Concepts approach is shown in Figure 2. Note that, Logistic Regression demonstrates the best performance among the selected classification algorithms based on its F1-score. The high value of the F1-score is mainly due to its strong recall. So, if we can improve the precision of this algorithm, then F1-score can also improve.

The Offline Translation performance shown in Figure 3 demonstrates uneven performance between the selected classification algorithms. For example, the Naive Bayes classifier has the worst F1-score among all the proposed approaches due to its very low recall. SMO is the best classification algorithm for the Online Translation approach as can be seen in Figure 4.

We compare all three approaches in Figure 5 based on the best classification algorithm for each approach. Although the Online Translation algorithm has the best performance overall, but the Multilingual Concepts approach demonstrates promising results with less computations, which necessitates further analysis.

IV. Demonstration

A screenshot of the demonstration is provided in Figure 1. It shows the data not only from Twitter, but also from both physical sensors and additional social networks related to landslide events on a web map. The data from physical sensors

---

3http://www.usgs.gov/
<table>
<thead>
<tr>
<th>Approach</th>
<th>Randomized ESA</th>
<th>Training Dataset</th>
<th>Evaluation Dataset</th>
</tr>
</thead>
<tbody>
<tr>
<td>Multilingual Concepts</td>
<td>En, Pt</td>
<td>En</td>
<td>Pt</td>
</tr>
<tr>
<td>Offline Translation</td>
<td>Pt</td>
<td>En → Pt</td>
<td>Pt</td>
</tr>
<tr>
<td>Online Translation</td>
<td>En</td>
<td>En</td>
<td>Pt → En</td>
</tr>
</tbody>
</table>

**TABLE II**
OVERVIEW OF PROPOSED APPROACHES

<table>
<thead>
<tr>
<th>Wikipedia articles</th>
<th>Training dataset</th>
<th>Evaluation dataset</th>
</tr>
</thead>
<tbody>
<tr>
<td>English</td>
<td>2,400</td>
<td>3,689</td>
</tr>
<tr>
<td>Portuguese</td>
<td>N/A</td>
<td>3,241</td>
</tr>
</tbody>
</table>

**TABLE III**
OVERVIEW OF DATASET

<table>
<thead>
<tr>
<th>Tweet in Portuguese</th>
<th>Meaning in English</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chuva provoca deslizamento de terra em Carapicuíba, na Grande SP.</td>
<td>Rain causes landslide in Carapicuíba, São Paulo metropolitan area.</td>
</tr>
<tr>
<td>Após deslizamento de encosta, mulher morre soterrada em Salvador.</td>
<td>Woman dies buried after hillside sliding in Salvador.</td>
</tr>
<tr>
<td>I7 pontos para entender o desabamento das Bolsas na China.</td>
<td>17 points to understand the stock market collapse in China.</td>
</tr>
<tr>
<td>Landslide não sai da minha cabeça, até quando eu tenho que estudar o.o</td>
<td>I can’t get Landslide song out of my mind, even when I need to study.</td>
</tr>
</tbody>
</table>

**TABLE IV**
EXAMPLES OF LANDSLIDE TWEETS IN PORTUGUESE

**Fig. 2.** Classification performance of Multilingual Concepts approach

**Fig. 3.** Classification performance of Offline Translation approach

**Fig. 4.** Classification performance of Online Translation approach

**Fig. 5.** Comparison of proposed approaches
include live feeds on earthquakes from USGS\textsuperscript{5} and on heavy rainfalls from TRMM\textsuperscript{6}. There is also a static feed of the global landslide hazard distribution.

The data from social networks include live feeds from Twitter, YouTube and Facebook. Users need to select a language that will filter the feeds from social networks by that language, such as Portuguese.

Users can view detailed information about items from each feed, e.g. they can read the contents of a tweet or watch a video from YouTube. They can also see all of the items that were used to make a decision regarding a landslide for each detected event provided in the Events feed.

Finally, users can toggle the checkbox for any feed to view the data from a particular feed or a combination of feeds.

V. CONCLUSION AND FUTURE WORK

In this paper we propose and evaluate three approaches for a multilingual automated classification of texts in any language without the need for a training dataset in that language. We use multilingual Wikipedia articles to build ESA matrices that represent the meaning of any text as a weighted vector of Wikipedia-based concepts in different languages. With this approach, it is sufficient to have a single annotated training dataset in a language of choice to automatically support classification of texts in other languages. We successfully apply the proposed approaches for disambiguation of landslide data from Social Media in English and Portuguese and display the results on a web map.

Our future work includes a more extensive evaluation of the proposed approaches based on larger datasets, and we will apply them to other problem domains and additional languages, such as Japanese and Chinese.

VI. ACKNOWLEDGMENTS

This research has been partially funded by National Science Foundation by CISE’s SAVI/RCN (1402266, 1550379), CNS (1421561), CRISP (1541074), SaTC (1564097) programs, an REU supplement (1545173), and gifts, grants, or contracts from Fujitsu, HP, Intel, and Georgia Tech Foundation through the John P. Imlay, Jr. Chair endowment. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation or other funding agencies and companies mentioned above.

REFERENCES


\textsuperscript{5}http://earthquake.usgs.gov/earthquakes/feed/v1.0/geojson.php
\textsuperscript{6}https://trmm.gsfc.nasa.gov/trmm_rain/Events/latest_3_day_landslide.html