Unsupervised Document Classification in Low-resource Languages for Emergency Situations

Anonymous ACL submission

1 Introduction

During emergency, relief workers need to constantly track updates so that they can learn of situations that require immediate attention (Stowe et al., 2016). However, it is challenging to carry out these efforts rapidly when the information is expressed in people’s native languages, which have little to no resources for NLP. We aim for building an adaptable language-agnostic system for such emergent situations that can classify incident language documents into a set of relevant fine-grained classes of humanitarian-needs and unrest situations. Our approach requires no language specific feature engineering and rather leverages the semantic difference between generic class features to build a classification framework that supports relief efforts. We assume no knowledge of the incident language, except the commonly available bilingual dictionaries (which tend to be very small or are generated from out-of-domain data such as Bible alignments). First, we obtain keywords for each target class using English news corpora (Naik et al., 2017; Marujo et al., 2015; Wen and Rosé, 2012; Özgür et al., 2005; Tran et al., 2013), that are then translated using the available bilingual dictionary (Zhang et al., 2016; Adams et al., 2017). Second, an unsupervised bootstrapping module enhances the generic keywords by adding incident-specific language-specific keywords (Knopp, 2011; Huang and Riloff, 2013; Ebrahimi et al., 2016). Next, we use all the keywords to generate labeled data. Finally, this data is used to train a downstream document classifier. This entire procedure is language-agnostic because it bypasses the necessity to create training data from scratch. We validate this procedure in a low-resource setup, with 7 distinct languages, showing significant improvements over the baseline by at least 13 F1 points. To the best of our knowledge, our approach is the first to combine the use of distant supervision from English and in-language semantic bootstrapping for such a low-resource task. We believe our method can form a strong benchmark for future developments in rapid low-resource unsupervised classification.

2 Approach

Figure 1 shows the overall architecture of our approach, composed of three primary modules. **Keywords**: We use English in-domain corpora viz. Google News and Relief-Web corpus\(^1\) to generate task-specific keywords, such that each keyword is strongly indicative of the underlying class(es) of a document. We cluster the documents based on their classes and use tf-idf to pick top 100 candidate words for each class. We then compute a label affinity score between each candidate and class labels using cosine-similarity between their corresponding Word2Vec embeddings\(^2\). In this way, each candidate keyword has a different association strength across all classes, and we only retain the ones above threshold 0.9. The pruned keywords are translated into the incident language using the available bilingual dictionary, dropping the ones that are absent. Finally, we use keyword spotting to label each document with class(es) of the keyword/s present in it. **Bootstrapping**: Dropping keywords during translation leaves a significant fraction of documents unlabeled. To improve the percentage of labeled documents, we expand the keywords within the incident language using a two-step process. First, we cluster the labeled documents (obtained from keyword spotting) based on their classes. For each word in a cluster, we compute the sum of its tf-idf score across other clusters and its average word

\(^1\)Pre-classified English documents into disaster relief needs and emergency situations (https://reliefweb.int)

\(^2\)https://code.google.com/archive/p/word2vec/
similarity with all other keywords present in that cluster. Each keyword belongs to the same class as its cluster. Second, we prune words with less than 0.9 score, and use the rest to again label more documents using keyword spotting (e.g. fraction of labeled documents in Uzbeck increased by 36%).

**Classification**: Finally, we use all labelled documents obtained from Keywords and Bootstrapping module to train a classifier, which classifies all the remaining incident language documents.

### 3 Experiments and Results

We used the LDC corpora\(^3\) for 7 low-resource languages having 11 class labels: *Crime-violence, Terrorism, Regime-change, Medical, Food, Water, Evacuation, Shelter, Search-rescue, Infrastructure,* and *Utilities*. Mandarin, Uzbek, Farsi and Spanish have 190 documents with average 2.7 labels per document. Tigrinya, Uyghur and Oromo have 1.1K, 3.6K and 2.7K documents with 1.4, 0.1 and 1.0 labels per document respectively. Apart from the difference in language families and writing scripts, morphological complexity adds further challenge to classification. As shown in Table 1\(^4\), the Keywords module results in highest average F1 gain over the baseline, showing the effectiveness of using language-agnostic information for tasks. As expected, this module primarily improves precision. Further, on average 84.53% keywords were dropped in translation, suggesting improvements in bilingual dictionary can benefit this module. Similarly, the Bootstrap module focuses on incident-specific information and primarily improves recall, resulting in an overall F1 gain of 6%. Finally, the Classifier module achieves an overall improvement of 4% F1. We observe low performance on Oromo, which is a morphologically rich language. On finer inspection, we found the corpus had several misspelled words. For instance, we identified different versions of *Ethiopia*, such as *itiyoophiyaa*. We also observe that the languages of same family like Uyghur and Uzbek have similar performances. In most cases, the gain in F1 provided by keyword extraction and bootstrapping is significantly higher than that from any classifier. This suggests that the classifier performance will improve only when we improve the mappings between source and target languages.

---

Table 1, the Keywords module results in highest average F1 gain over the baseline, showing the effectiveness of using language-agnostic information for tasks. As expected, this module primarily improves precision. Further, on average 84.53% keywords were dropped in translation, suggesting improvements in bilingual dictionary can benefit this module. Similarly, the Bootstrap module focuses on incident-specific information and primarily improves recall, resulting in an overall F1 gain of 6%. Finally, the Classifier module achieves an overall improvement of 4% F1. We observe low performance on Oromo, which is a morphologically rich language. On finer inspection, we found the corpus had several misspelled words. For instance, we identified different versions of *Ethiopia*, such as *itiyoophiyaa*. We also observe that the languages of same family like Uyghur and Uzbek have similar performances. In most cases, the gain in F1 provided by keyword extraction and bootstrapping is significantly higher than that from any classifier. This suggests that the classifier performance will improve only when we improve the mappings between source and target languages.

---

\(\text{\textsuperscript{3}}\)https://www.ldc.upenn.edu

\(\text{\textsuperscript{4}}\)We use the LOREHLT evaluation guidelines (https://goo.gl/ZT7sMq) for scoring
References


