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 atleast 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/

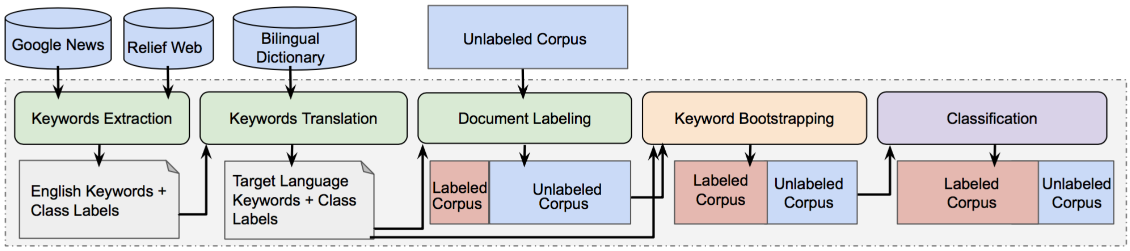

Figure 1: System architecture for low-resource document classification.

	Mandarin			Spanish			Uzbek			Farsi			Tigrinya			Uyghur			Oromo		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1
Random	0.09	0.09	0.09	0.09	0.09	0.09	0.09	0.09	0.09	0.09	0.09	0.09	0.09	0.09	0.09	0.09	0.09	0.09	0.09	0.09	0.09
+ KWD	0.40	0.22	0.28	0.57	0.14	0.23	0.51	0.11	0.18	0.49	0.18	0.26	0.54	0.55	0.55	0.61	0.28	0.39	0.45	0.08	0.14
+ BS	0.34	0.38	0.36	0.22	0.74	0.34	0.33	0.47	0.39	0.31	0.30	0.30	0.53	0.59	0.56	0.52	0.31	0.39	0.33	0.09	0.14
+ KNN	0.31	0.40	0.35	0.31	0.46	**0.37**	0.29	0.47	0.36	0.24	0.48	0.32	0.54	0.58	0.56	0.38	0.36	0.37	0.10	0.12	0.11
SVM	0.30	0.39	0.34	0.30	0.45	0.36	0.34	0.48	**0.40**	0.29	0.29	0.29	0.55	0.58	0.57	0.47	0.38	0.41	0.20	0.09	0.12
R-Forest	0.37	0.38	**0.38**	0.31	0.46	**0.37**	0.33	0.47	0.39	0.23	0.40	0.30	0.56	0.61	0.58	0.42	0.38	0.40	0.12	0.10	0.11
Log-Reg	0.29	0.39	0.33	0.31	0.45	**0.37**	0.32	0.47	0.38	0.23	0.37	0.29	0.56	0.61	0.58	0.51	0.36	**0.42**	0.12	0.10	0.10
GNB	0.32	0.40	0.36	0.30	0.46	0.36	0.33	0.51	**0.40**	0.28	0.31	0.30	0.56	0.62	**0.58**	0.47	0.37	0.41	0.11	0.10	0.11
DAN	0.22	0.38	0.27	0.21	0.74	0.33	0.23	0.52	0.32	0.25	0.68	**0.37**	0.56	0.58	0.57	0.37	0.38	0.37	0.26	0.19	**0.22**
LSTM	0.29	0.39	0.33	0.43	0.23	0.30	0.29	0.48	0.37	0.25	0.24	0.24	0.51	0.63	0.56	0.37	0.31	0.33	0.09	0.20	0.13

Table 1: Results of classification across 7 languages, over each module (Modules - KWD:Keywords, BS:Bootstrap; Classifiers - R-Forest:Random Forest, GNB:Gaussian Naive Bayes, Log-Reg:Logistic regression, LSTM (Hochreiter and Schmidhuber, 1997) and DAN (Iyyer et al., 2015; Chen et al., 2016)

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.

[3]https://www.ldc.upenn.edu

[4]We use the LOREHLT evaluation guidelines (https://goo.gl/ZT7sMq) for scoring

References

Oliver Adams, Adam Makarucha, Graham Neubig, Steven Bird, and Trevor Cohn. 2017. Cross-lingual word embeddings for low-resource language modeling. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, volume 1, pages 937–947.

Xilun Chen, Yu Sun, Ben Athiwaratkun, Claire Cardie, and Kilian Weinberger. 2016. Adversarial deep averaging networks for cross-lingual sentiment classification. *arXiv preprint arXiv:1606.01614*.

Javid Ebrahimi, Dejing Dou, and Daniel Lowd. 2016. Weakly supervised tweet stance classification by relational bootstrapping. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1012–1017. Association for Computational Linguistics.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Ruihong Huang and Ellen Riloff. 2013. Multi-faceted event recognition with bootstrapped dictionaries. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 41–51.

Mohit Iyyer, Varun Manjunatha, Jordan Boyd-Graber, and Hal Daumé III. 2015. Deep unordered composition rivals syntactic methods for text classification. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, volume 1, pages 1681–1691.

Johannes Knopp. 2011. Extending a multilingual lexical resource by bootstrapping named entity classification using wikipedia's category system. In *Proceedings of the Fifth International Workshop On Cross Lingual Information Access*, pages 35–43. Asian Federation of Natural Language Processing.

Luis Marujo, Wang Ling, Isabel Trancoso, Chris Dyer, Alan W Black, Anatole Gershman, David Martins de Matos, João Neto, and Jaime Carbonell. 2015. Automatic keyword extraction on twitter. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 637–643. Association for Computational Linguistics.

Aakanksha Naik, Chris Bogart, and Carolyn Rose. 2017. Extracting personal medical events for user timeline construction using minimal supervision. *BioNLP 2017*, pages 356–364.

Arzucan Özgür, Levent Özgür, and Tunga Güngör. 2005. Text categorization with class-based and corpus-based keyword selection. In *International Symposium on Computer and Information Sciences*, pages 606–615. Springer.

Kevin Stowe, Michael J. Paul, Martha Palmer, Leysia Palen, and Kenneth Anderson. 2016. Identifying and categorizing disaster-related tweets. In *Proceedings of The Fourth International Workshop on Natural Language Processing for Social Media*, pages 1–6. Association for Computational Linguistics.

Dang Tran, Cuong Chu, Son Pham, and Minh Nguyen. 2013. Learning based approaches for vietnamese question classification using keywords extraction from the web. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, pages 740–746.

Miaomiao Wen and Carolyn Penstein Rosé. 2012. Understanding participant behavior trajectories in online health support groups using automatic extraction methods. In *Proceedings of the 17th ACM international conference on Supporting group work*, pages 179–188. ACM.

Dongxu Zhang, Boliang Zhang, Xiaoman Pan, Xiaocheng Feng, Heng Ji, and XU Weiran. 2016. Bitext name tagging for cross-lingual entity annotation projection. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 461–470.