Automatic Extraction of Locations from News Articles Using Domain Knowledge

Loitongbam Sanayai Meetei(✉), Ringki Das, Thoudam Doren Singh, and Sivaji Bandyopadhyay

Department of Computer Science and Engineering, National Institute of Technology Silchar, Silchar, Assam, India
loisanayai@gmail.com, ringkidas@gmail.com, thoudam.doren@gmail.com, sivaji.cse.ju@gmail.com

Abstract. With the increasing amount of digital data, it is becoming increasingly hard to extract useful information from text data, especially for resource-constrained languages. In this work, we report the task of language-independent automatic extraction of locations from news articles using domain knowledge. The work is tested on four languages namely, English and three resource-constrained languages: Assamese, Manipuri and Mizo, the lingua francas of three neighboring North-Eastern states of India namely Assam, Manipur, and Mizoram respectively. Our architecture is based on semantic similarity between similar words based on the popular word embedding, word2vec model coupled with the domain knowledge of the aforementioned regions. The model is able to detect the best possible detailed locations.

Keywords: Location extraction · Word2vec · Word similarity · Resource constrained language · Assamese · Manipuri · Mizo

1 Introduction

The North-Eastern part of India (officially North Eastern Region) comprises of seven sister states with around 220 spoken languages. Assamese, Manipuri and Mizo, the lingua francas of three neighboring states: Assam, Manipur, and Mizoram respectively, belong to different language families. Assamese is the major language spoken in the North-Eastern part of India and an official language of Assam. It belongs to the Indo-European family of languages. Manipuri comes under the Tibeto-Burman language while Mizo is a Kuki-Chin language. Both Tibeto-Burman and Kuki-Chin are the subfamily of the Sino Tibetan language. Assamese is spoken by over 15 million speakers mainly in the Brahmaputra valley. The speaker of Manipuri and Mizo are around 3 million and 1 million respectively. The usage of all the three languages can be found in the Northeastern states as well as in the neighboring countries like Bangladesh and Myanmar. There are six major types of word order made up of three building blocks, namely Subject(S), Object(O) and Verb(V), namely: i) SVO, ii) SOV, iii) VSO iv) OSV v) VOS and vi) OVS. The word order of Assamese, English, Manipuri, and Mizo are SOV, SVO, SOV,

R. Patgiri et al. (Eds.): BigDML 2019, CCIS 1317, pp. 36–47, 2020.
https://doi.org/10.1007/978-3-030-62625-9_4

and OSV respectively. Even if Assamese and Manipuri belong to a different language family group, both the languages follow the SOV pattern. In some cases, the word order in the Assamese tends to be relatively free. Each of the three states is divided into 33, 9[1] and 8 districts in Assam, Manipur, and Mizoram respectively.

Named Entity Recognition (NER) is one of the popular tasks in natural language processing (NLP). The idea here is to identify and extract the named entities such as the names of persons, locations, organizations, etc. from a text. The state of the art of the NER system for a widely used language such as Chinese, English, etc. are close to the human brain. A quality NER system requires an annotated training corpus, however, for low resource-constrained languages, preparing such training corpus is a laborious task. Language-specific tools such as POS tagger which helps in enhancing the performance of the NER system is also not available for such languages.

In this paper, we proposed a language-independent system for the automatic extraction of locations from local daily news articles using domain knowledge. Here, the domain knowledge refers to a list of pre-defined names of the location in a region. Our system doesn't use language tools such as Part of Speech (POS) tagger and is based on word embedding and word similarity measure. Also, the system is able to identify the locations at the possible detailed location, such as at the locality level. The system can be used in a news recommendation system, disaster management, etc.

The rest of this paper is divided into four sections, with Sect. 2 discussing the related previous works. Section 3 and Sect. 4 describe the architecture used in the model and the analysis of experimental results respectively. Finally, summing up with the conclusion and future work in Sect. 5.

2 Literature Review

In the last decennium, word association has been well studied in linguistics. "Word Association" as a psycholinguistic term proposed by Church et al. [3] e.g. co-occurrence with other words, has a variety of applications which can be used in optical character recognition (OCR), speech recognition, information retrieval and enhancing the productivity of lexicographers.

Previous research works have carried out the study on the semantic similarity of text on different levels, such as word, sentence, and phrase, etc. The proposed model by Mihalcea et al. [9] was to find out the semantic similarity of short texts using Corpus and Knowledge-based. Along with the vector-based similarity approach, Microsoft paraphrase parallel corpus was used for the identification of paraphrase as well as to generate paraphrase. Islam et al. [6] proposed a semantic text similarity model (STS) for measuring similar text from semantic and syntactic information based on similar functions. A ranked-based system was generated by Wen et al. [16] for identifying synonyms which are based on the distributional hypothesis. Sentence similarity can be applied in many areas such as text mining, question-answering, and text summarizing according to Achananuparp et al. [1]. Wang et al. [15] developed a model to estimate sentence similarity with the help of the decomposition-composition vector and CNN (Convolution Neural Network) model.

[1] 16 districts as of December 2016, demarcation yet to be ascertained.

From a large dataset, two atypic models were suggested by Mikolov et al. [10] for computing continuous vector representation of words. It was remarked that at lower cost learning high-quality word vectors is better than neural networks. Skip-gram and CBOW (Continuous Bag of Words) can give better results than Neural Network Language Model and Recurrent Neural Network Language Model. A negative sampling is reported by Goldberg et al. [4] as an efficient approach to word embedding for finding out words having a similar meaning with similar context. It brings a different objective to the Skip-gram model.

The extraction of geolocation from text data was put forward for consideration by Imani et al. [5]. Named Entity Recognition tool has been used to identify geolocation from news articles using supervised classification and sentence embedding techniques at the country level. With the help of Named Entity Recognizer and geo-tagged tweets, the disaster location from Microblogs was extracted in Lingad et al. [7]. The affected area could be located as geographic locations like country, city, village, and river, etc. and point-of-interest could be hotels, shopping centers, and restaurants. Different tools such as Twitter, Open NLP, Standford NER, Yahoo! Place Maker for recognizing references to the location were used.

Named Entity Recognition (NER) is considered to be a baby step in the direction of Information Extraction which also plays a very vital role in Natural Language Processing (NLP). Existing works on NER for regional languages were carried out by [2, 12, 14]. Named Entity Recognition for the Assamese language was reported by Sharma et al. [12] using ruled based and supervised machine learning techniques such as Hidden Markov Models (HMM), Conditional Random Fields (CRF), Support Vector Machines (SVM) and Maximum Entropy (ME). Because of the scarcity of resources in a regional language, it is a very challenging task for their experiments. Recognition of named entity in Manipuri is reported by Singh et al. [14]. The authors developed two different models using an active learning technique and Support Vector Machine. A large amount of annotated corpora was used in both the models to recognize the location, person and organization names, designations, etc. For Mizo language a NER system is expressed by Bentham et al. [2] using a rule-based approach on news corpus. A Named Entity Extraction system was proposed by Nadeau et al. [11] using unsupervised techniques. To identify and classify entities from a given document the authors have to combine the Named Entity Extraction with Named Entity Disambiguation. The authors created a large cluster of semantically related words using seed words.

3 Architecture

A high-level graphical representation of the proposed schema is introduced in Fig. 1. The first step is a collection of data from different local daily newspapers in four different languages, namely Assamese, English, Manipuri, and Mizo. The data collection step is followed by the data cleaning process. Using the processed data, a word embedding model is trained using the word2vec. Finally, retrieving the candidate keywords based on the ranking of the cosine similarity distance with the seed words prepared with the help of domain knowledge. The approach of our model is described in the following sub-sections.

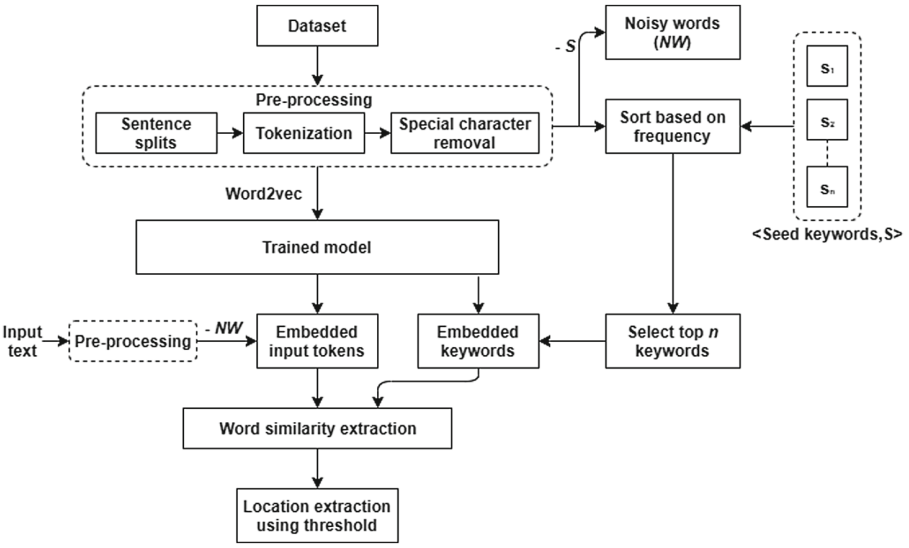

Fig. 1. Our proposed model

3.1 Data Collection

We prepared our dataset by collecting the news article from the local daily newspapers separately for Assamese[2], Manipuri[3] and Mizo[4] languages. Apart from this dataset, we also collected the news articles from the local daily newspaper of Manipur[5] reported in the English language. For Assamese and English languages, separate scrappers built inhouse using Java and its Jsoup library[6] is used. As for the Manipuri and Mizo language, we have used the dataset in [13] and [8] respectively. The news articles on the Assamese language are collected for the period of June 2018 to June 2019. While for English, Manipuri and Mizo[7] are collected for the period of July 2011 to May 2019, May 2008 to May 2010 and April 2013 to June 2019 respectively. The collected dataset is used as a training dataset for our model.

3.2 Pre-processing

Data collected from the news articles are in an unstructured format and contains several noisy texts. To process the data into a consumable format, the whole corpus grouped by language is subjected to a series of pre-processing steps separately.

The pre-processing step includes:

[2] https://www.asomiyapratidin.in/.
[3] https://www.thesangaiexpress.com/.
[4] https://www.vanglaini.org/.
[5] https://www.ifp.co.in/.
[6] https://jsoup.org/.
[7] We collected more dataset on [8].

1. Splitting the text into sentences.
2. Tokenization of a sentence into token of words.
3. Removing special characters and punctuation.
4. Replacing multiple spaces into a single space.
5. Removal of single-character tokens.
6. Normalize to lower case [only for Roman characters].

The result is a list of sentences for each of the languages. The details of the training corpus (number of sentences, number of tokens and the average number of tokens per sentence for each language) after the completion of the pre-processing are shown in Table 1. What can be clearly seen in Table 1 is that the dataset for Mizo language is the largest, followed by English, Manipuri and Assamese.

Table 1. Statistics of our training corpus of each of the languages.

Language	Sentences	Tokens	Avg tokens per sentence
Assamese	69416	839062	12
English	660444	17732538	26
Manipuri	104625	1988554	19
Mizo	746609	22431244	30

3.3 Word2vec

Word embedding is a vector representation of the text vocabulary, capable of highlighting the semantic or syntactic similarities or relations between words. One of the most popular technique to learn word embedding is word2vec, a shallow neural network model. The word2vec model consists of three main layers namely, an input layer, a hidden layer, and an output layer as shown in Fig. 2. The input layer is a text corpus with C context words where each word is represented by the one-hot encoding vector. The hidden layer is a fully-connected layer with n neurons whose weights are the word embedding. The output layer is a feature vector of length V. It groups the similar word vector into one vector space. The word2vec can be trained with either of the two language models: CBOW or Skip-gram. For building the word embedding, we used Gensim[8] which is a Python library for extracting document with similar semantic.

3.4 Feature Selection

The feature selection of our model consists of two main parts, namely, filtering of keywords and preparing a list of noisy words. Our main feature is the use of custom seed keywords after filtering based on domain knowledge of each region, namely Assam

[8] https://radimrehurek.com/gensim/.

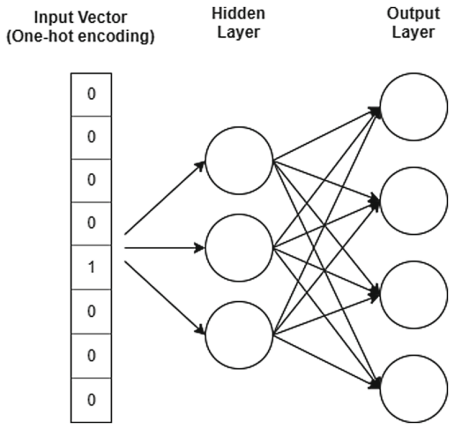

Fig. 2. Word2vec model

(news report in Assamese), Manipur (news report in English and Manipuri) and Mizoram (news report in Mizo). The following steps are applied dataset to a training to generate the final seed keywords:

1. Prepare list of keywords ($D = d_1, d_2 \ldots d_p$) with names of popular locations of a region.
2. For each unique tokens ($T = t_1, t_2 \ldots t_n$) in the training dataset, prepared their corresponding frequency list ($F = f_1, f_2 \ldots f_n$) in the dataset.
3. Rank the keywords (D) based on their number of occurrences.
4. Select the top m highest frequency keywords, which is our final seed keywords ($K = k_1, k_2 \ldots k_m$) . The value of m range from 5 to 10.

Further, a list of top N highest frequency tokens are selected from F. However, we exclude any tokens present in D from the list. The value of N range from 100 to 150. Using these high-frequency words, we build a list of "noisy-words" (NW) for each dataset as these words tend to be comprised of non-noun words such determiner, preposition, etc. which are irrelevant to the candidate keywords for locations.

3.5 Cosine Similarity

For a large document, only counting the maximum common words for similarity measurement sometimes will not work properly. Cosine similarity helps overcome the traditional flaw. It is a well known metric to measure the similarity between different text. Cosine similarity calculates the cosine angle between two vectors to measure the similarity between them.

If $\vec{w}_1$ and $\vec{w}_2$ are two word vectors, then the cosine distance between them is calculated as:

$$cos(\theta) = \frac{\vec{w}_1 . \vec{w}_2}{||\vec{w}_1||\ ||\vec{w}_2||} \tag{1}$$

where dot (.) indicates dot product and $||\vec{w}_1||$ and $||\vec{w}_2||$ are the length of the vector $\vec{w}_1$ and $\vec{w}_2$ respectively.

Finally, the cosine similarity between $\vec{w}_1$ and $\vec{w}_2$ is computed as:

$$\cos(\vec{w}_1, \vec{w}_2) = 1 - cos(\theta) \tag{2}$$

We used Gensim[9] tools to compute the similarity between the words.

3.6 Steps for Location Extraction

The steps used for identifying the candidate terms in an input news article are as follows:

1. Remove the noisy-words from the input text after applying the pre-processing step describe in Sect. 3.2.
2. Generate all the unique tokens T from the result of Step 1.
3. Load the trained word2vec model for the language.
4. Compute the sum of cosine distance ($S = s_1, s_2 \ldots s_n$) for $t_1, t_2 \ldots t_n \in T$, with each of the seed keyword $k_1, k_2 \ldots k_m \in K$ such that:

$$s_j = \sum_{j=1}^{m} \mathrm{t}_i.k_j \tag{3}$$

5. Sort the tokens T based on the ranking of the cosine distance (S).
6. If $T <= 200$, select the top 10 from the sorted list else select the top 20.

The evaluation is carried out based on the number of unique location terms in the ground truth reference present on the list of terms (i.e. either in the top 10 or the top 20) generated by the system.

4 Experiment and Results

The dataset collected for each language is subjected to the same architecture described in Sect. 3 separately. The training dataset is used to build the word embedding model for each of the languages separately using word2vec. The word2vec is trained with different values of *window* and *size*. Here, the window represents the maximum distance between the current and predicted word within a sentence and size is the dimensionality of the word vector. During our experiment, the word2vec trained on the $Skip - gram$ language model with $window = 5$, $size = 300$ and $learning\ rate = 0.001$ is observed to be an optimal one. A sample of the word2vec model trained on the dataset of Mizo language is shown in Fig. 3. In the sample, we checked the 10 most similar words for the keyword "aizawl" (the capital city of Mizoram), and the generated words are found to be the location names in Mizoram. This result highlights that the vectors of the location are mapped in the same space.

[9] https://radimrehurek.com/gensim/.

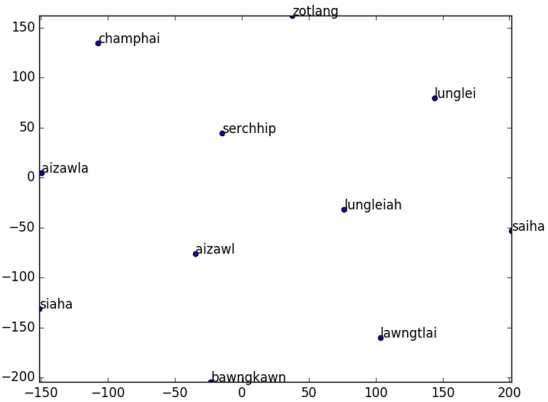

Fig. 3. Sample of trained word2vec model

After training the collected dataset using word2vec, we finally get four word embedding models for each of the languages. We evaluate our proposed model using a test dataset comprising of 20 news article items each on Assamese, English, Manipuri, and Mizo languages separately from the same source as mentioned in Sect. 3.1. The location in each of the news articles from the test dataset is manually tagged by a native or a fluent speaker who also have the domain knowledge about the regions. The annotated dataset is used as our ground truth reference. We employed the steps detailed in Sect. 3 on the test dataset to generate the output. A detailed evaluation of our model is summarized in the Table 2. Table 2 shows the languages followed by the total number of news articles and the combined total number of words in the test dataset, the number of seed keywords used to identify the locations in the input text, the total number of location terms present in the input text (LP), the total number of location terms detected by our model in the test dataset (LD), and recall. The values of recall in Table 2 are in terms of the percentage and is computed as:

$$\text{Recall} = \frac{LD}{LP} \times 100 \tag{4}$$

As shown in Table 2, the model for Mizo language achieve highest accuracy of 92.3% and the model for Assamese language achieve lowest accuracy of 29.7%.

Table 2. Statistics of the output obtained from the model

Language	Articles-words	# Keywords	LP	LD	Recall
Assamese	20–2415	10	47	14	29.7
English	20–5374	7	82	64	78.04
Manipuri	20–6753	7	76	24	31.58
Mizo	20–2312	8	39	36	**92.3**

4.1 Sample Input-Output

A sample input and output for each of the languages is listed below as: the input text in the language, translation to English of the input text, ground truth terms in the input text, and finally the output produced by our system.

Language: Assamese

Input: চৰকাৰী ভূমি বেদখলৰ যেন প্ৰতিযোগিতা চলিছে বটদ্ৰৱাত। বটদ্ৰৱাৰ ভোমোৰাগুৰি গ্ৰেজিং ৰিজাৰ্ভ, বটদ্ৰৱা চৰকাৰী বীজ পাম, শ্ৰীমন্ত শংকৰদেৱ বিশ্ববিদ্যালয়ৰ আৱন্টিত ভূমিৰ কাষৰ চৰকাৰী ভূমি আদি বিভিন্ন চৰকাৰী ভূমি বেদখলৰ বাতৰি প্ৰকাশ পাই থকাৰ সময়তে বটদ্ৰৱা বিধানসভা সমষ্টিৰ অন্তৰ্গত শিলপুখুৰীত চৰকাৰী ভূমি বেদখলৰ অভিযোগ পোৱা গৈছে। প্ৰাপ্ত অভিযোগ অনুসৰি, মৰিগাঁও জিলাৰ মিকিৰভেটা ৰাজহ চক্ৰৰ অন্তৰ্গত বড়িবজা-ৰস্থিত শিলপুখুৰী গাওঁ পঞ্চায়তৰ কাৰ্যালয়ৰ সন্মুখত থকা চৰকাৰী ভূমি যোৱা কিছুদিন ধৰি নিশাৰ ভিতৰতে বেদখল কৰিছে একাংশ দুষ্ট প্ৰকৃতিৰ লোকে।

(*Translation: It seems like a competition is going on to possess govt lands ille- gally at Botodrowa. Amidst the circulation of news on illegal land possession near areas like Botodrowa's Bhomoraguri Grazing Reserve, Botodrowa Government Seed Palm, Srimanta Shankardev University, complaints have been found that there is an illegal possession of government lands at Silpukhuri under Botodrowa legislative assembly. According to the complaint, govt land in front of Silpukhuri Gaon Panchayat, which is situated at Baribazaar area of Marigaon district's Mikirveta Tax Circle, has been illegally possessed by a few miscreants in a matter of days.*)

Ground truth terms in input text: বটদ্ৰৱা, শিলপুখুৰী, মৰিগাঁও
Output: শংকৰদেৱ, অন, মৰিগাঁও, ৰাজহ, নিশাৰ, চক, শিলপুখুৰী, ভূমিৰ, বীজ, টিত

Language: English

Input: A family was poisoned after consuming wild mushrooms. Three persons from Sihai Khullen of Ukhrul district passed away after consuming a mushroom which is locally known as "Ngatha Var." The deceased includes a 70 year old man and his two grandchildren. According to information culled from the villagers, the deceased persons have been identified as Joseph Khangrah, 70; Puimi Khangrah, 5, daughter of Vareiso Khangrah; and Chalakmi Khangrah, 3, son of Vareiso Khangrah. Vareiso Khangrah, father of two deceased children and his wife, Timnah Khangrah along with one of his daughters, Wongayung Khangrah were admitted for medical treatment at RIMS along with some of his relatives who also tasted the dish. They are out of danger and have been discharged from RIMS. As per information made available, the family prepared the mushroom dish on July 2 last for their dinner. After consuming, they vomited but they did not seek medical treatment thinking it would be okay. Joseph Khangrah died on the morning of July 8 after complaining of stomach problems. Chalakmi Khangrah died on the evening of July 10 while Puimila Khangrah died on the way to Imphal after crossing Lambui at around 9:00 pm of July11. She was advised and referred to go to Imphal by doctors at Ukhrul District Hospital. It may be mentioned here that Sihai Khullen is situated about 37 km from Ukhrul headquarter; and the village falls under Khangkhui Primary Health Centre (PHC), which is much farther away from Ukhrul District Hospital.

The connectivity road is in a pathetic condition which has caused villagers to face a lot of hardships especially during rainy season and availing basic medical treatment.
Ground truth terms in input text: Sihai Khullen, Ukhrul, Imphal, Lambui **Output:** ukhrul, headquarter, district, lambui, khangkhui, khullen, sihai, situated, imphal, phc
Language: Manipuri

Input:তমেংলোং ডিষ্ট্রিক্ট কন্ট্রেক্টরস এসোসিয়েসনগী প্রসিডেন্ট ওইবা স্পিসিয়েল কন্ট্রেক্টর কলানচুং কামৈবু এনএসসিএন-আইএম, জেলিয়াংরোং রিজনগী কাংবুনা তমথীবা মওংদা ওত্ নৈরকপা থৌওংবু এসোসিয়েসন অসিনা অকনবা মওংদা কণ্ডেম তৌরক্লি| ঙরাং হেলঠ মিনিষ্টর পিঐচ পরিজাত্কি তমেংলোং খোঙচত্ মনুংদা ডিষ্ট্রিক্ট হোস্পিটাল, তমেংলোংগী কমপ্লেক্সতা মীটিং চত্থরিঙৈ নুংথিল পুং 2.30 রোম তাবদা কলানচুং কামৈবু এনএসসিএন-আইএমগী জেলিয়াংরোং রিজনগী সিইও লোংচাইবি গোনমৈনা ৱারী অমা শান্নবা পাম্মী হায়-দুনা লুপ অসিগী কাডর অমনা কৌথোকখিবনি হায়রি | কলানচুং কামৈবু জিপসি অমগী মনুংদা নমশিনখি অমসুং সিএওগী ওফিসতা পুখি অদুগা সিএও জেলিয়াংরোং রিজন এনএস-সিএন আইএমনা কৌবদা লাক্তে অমসুং নবেম্বর মঙাদা হেলঠ মিনিষ্টরগী ওফিস চেম্বরদা পাঙথোকখিবা মীটিং অমদা এনএসসিএন-আইএম জেলিয়াংরোং রিজনবু ইকায়বা পীখি হায়না মরাল শীরকপদা মহাক্না য়াদবগী মমিত্-মমায় পুনশিল্লগা পুং অমা মখায়রোম মীনুংশি হৈতনা চৈনা পংফুদা ফুদুনা ওত্ নৈরকখিবনি হায়রি | অসিগুম্বা থৌওং অসি মতুংদা অমুক হন্না চত্থদনবসু এসোসিয়েসন অসিনা মরী লৈনবা পুম্নমক্তা আপিল তৌরক্লি|

(*Translation: Tamenglong district contractor's association condemned the ha rassment and illtreatment of its president, special contractor Kalanchung Kamei by NSCN (IM) Jeliangrong region. Yesterday, during the health minister PH. PARIJAT tamenglong visit and an ongoing meeting inside the district hospital at around 2:30 p.m; Kalanchung Kamei was called out by the cadre of NSCN (IM) as he wanted to talk by its (C.O.) Longchaibi Gonmei. Kalanchung Kamei was dumped in a jipsy and carried away to the office of C.A.O, then he was blamed for not coming to the call of C.A.O. Jeliangrong region NSCN (IM) and also on 5th November in the office chamber of health minister. He make ashamed the NSCN (IM) Jeliangrong region, but he denied the charges, then after blindfolded he was beaten mercilessly for about half an hour. So, the association appeal to all not to repeat such acts in the future.*)

Ground truth terms in input text: তমেংলোং, জেলিয়াংরোং রিজন
Output: তমেংলোং, কাংবুনা, মঙাদা, ওফিস, জেলিয়াংরোং, পুখি, জিপসি, সিইও, রিজনগী, রিজন

Language: Mizo
Input: Mizoram Treasury Accounts Service Association (MITASA) chuan nimin khan Hotel Regencyah inkhawmpui an hmang. Khawzawl, Hnahthial leh Saituala Treasury Office sorkarin hawn a tum chu lawmawm an tih thu an tarlang a, "Chutih rualin, Mizoram Treasury hi siamṭhat (restruct) a ngai a, hei hi remchangah sorkar hnenah thlen ni se," an ti.

(*Translation: Mizoram Treasury Accounts Service Association (MISTA) held their conference yesterday at hotel regency. The association announced that it was happy to hear that khawzawl, hnahthial and saitual treasury will be establish soon by the government. In the meantime, Mizoram Treasury needs to be reconstruct and submit it to the government as soon as possible.*)
Ground truth terms in input text: Khawzawl, Hnahthial, Saitual

Output: treasury, mizoram, khawzawl, hnahthial, saituala, hawn, regency, accounts, inkhawmpui, chutih.

The system is able to detect the candidate words of locations at the locality level. One of the limitations of our model is inability to differentiate the ambiguity of terms which is used not only as a location, but also as a person's name, organization name's, etc. The system is also not able to detect the exact total number of locations present in the input text.

5 Conclusion

Tools such as POS tagger helps in improving the application of NER tasks, as most of the entity tag belongs to certain parts of speech such as noun. For a low resource-constrained language, building quality language-specific tools such as the POS tagger is a laborious task as it requires an annotated training corpus. In our work, we build a language-independent automatic extraction of location from news articles using seed keywords from the domain knowledge. Our system doesn't rely on language tools such as POS tagger, but instead is purely based on word embedding and the word similarity measure. We tested the feasibility of our model on three low resource-constrained languages belonging to different language groups and also on the English language. Our model is observed to produce a better result on the English and Mizo language as compared to the other remaining language used in our experiment. From the experiment, we observed that the word embedding model trained on a larger dataset performs better than the one trained on a smaller dataset. For regional languages like Assamese and Manipuri, location extraction is a very strenuous and daring task and suffers from a scarcity of resources. Our system is not able to differentiate the ambiguity behavior where the same word is used as a person's name and also as a location's name. In the future, a more enhanced language-independent model that can detect multiple entities can be explored. The model can also be improved in such a way that it can be measured with better evaluation parameters.

Acknowledgement. The authors appreciate the anonymous reviewers for their valuable and profound comments. The authors wish to express their thanks to Elcy S. Lalropeki and L. Tamphangambi for their assistance in this work.

References

1. Achananuparp, P., Hu, X., Shen, X.: The evaluation of sentence similarity measures. In: Song, I.-Y., Eder, J., Nguyen, T.M. (eds.) DaWaK 2008. LNCS, vol. 5182, pp. 305–316. Springer, Heidelberg (2008). https://doi.org/10.1007/978-3-540-85836-2_29
2. Bentham, J., Pakray, P., Majumder, G., Lalbiaknia, S., Gelbukh, A.: Identification of rules for recognition of named entity classes in mizo language. In: Fifteenth Mexican International Conference on Artificial Intelligence (MICAI), pp. 8–13. IEEE (2016)
3. Church, K.W., Hanks, P.: Word association norms, mutual information, and lexicography. Comput. Linguist. **16**(1), 22–29 (1990)

4. Goldberg, Y., Levy, O.: word2vec explained: deriving Mikolov et al.'s negative- sampling word-embedding method. arXiv preprint arXiv:1402.3722 (2014)
5. Imani, M.B., Chandra, S., Ma, S., Khan, L., Thuraisingham, B.: Focus location extraction from political news reports with bias correction. In: IEEE International Conference on Big Data (Big Data), pp. 1956–1964. IEEE (2017)
6. Islam, A., Inkpen, D.: Semantic text similarity using corpus-based word similarity and string similarity. ACM Trans. Knowl. Discovery Data (TKDD) **2**(2), 10 (2008)
7. Lingad, J., Karimi, S., Yin, J.: Location extraction from disaster-related microblogs. In: Proceedings of the 22nd International Conference on World Wide Web, pp. 1017–1020. ACM (2013)
8. Meetei, L.S., Singh, T.D., Bandyopadhyay, S.: Extraction and Identification of Manipuri and Mizo Texts from Scene and Document Images. In: Deka, B., Maji, P., Mitra, S., Bhattacharyya, D.K., Bora, P.K., Pal, S.K. (eds.) PReMI 2019. LNCS, vol. 11941, pp. 405–414. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-34869-4_44
9. Mihalcea, R., Corley, C., Strapparava, C., et al.: Corpus-based and knowledge-based measures of text semantic similarity. In: AAAI, vol. 6, pp. 775–780 (2006)
10. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781 (2013)
11. Nadeau, D., Turney, P.D., Matwin, S.: Unsupervised named-entity recognition: generating gazetteers and resolving ambiguity. In: Lamontagne, L., Marchand, M. (eds.) AI 2006. LNCS (LNAI), vol. 4013, pp. 266–277. Springer, Heidelberg (2006). https://doi.org/10.1007/11766247_23
12. Sharma, P., Sharma, U., Kalita, J.: The first steps towards Assamese named entity recognition. In: Brisbane Convention Center, vol. 1, pp. 1–11 (2010)
13. Singh, T.D., Bandyopadhyay, S.: Web based Manipuri corpus for multiword ner and reduplicated MWEs identification using Svm. In: Proceedings of the 1st Workshop on South and Southeast Asian Natural Language Processing, pp. 35–42 (2010)
14. Singh, T.D., Nongmeikapam, K., Ekbal, A., Bandyopadhyay, S.: Named entity recognition for Manipuri using support vector machine. In: Proceedings of the 23rd Pacific Asia Conference on Language, Information and Computation, vol. 2, pp. 811–818 (2009)
15. Wang, Z., Mi, H., Ittycheriah, A.: Sentence similarity learning by lexical decomposition and composition. arXiv preprint arXiv:1602.07019 (2016)
16. Wen, Y., Yuan, H., Zhang, P.: Research on keyword extraction based on word2vec weighted textrank. In: 2nd IEEE International Conference on Computer and Communications (ICCC), pp. 2109–2113. IEEE (2016)