Overview of Arnekt IECSIL at FIRE-2018 Track on Information Extraction for Conversational Systems in Indian Languages

Barathi Ganesh H B[1,2], Soman KP[1], Reshma U[2] Mandar Kale[2], Prachi Mankame[2], Gouri Kulkarni[2], Anitha Kale[2], and Anand Kumar M[3]

[1] Center for Computational Engineering & Networking (CEN) ,
Amrita School of Engineering, Coimbatore
Amrita Vishwa Vidyapeetham, India
barathiganesh.hb@arnekt.com
[2] Arnekt Solutions Pvt. Ltd., Pune, Maharashtra, India, 411028.
reshma.u@arnekt.com
[3] Department of Information Technology,
National Institute of Technology Karnataka
Surathkal, Mangalore.

Abstract. This overview paper describes the first shared task on Information Extractor for Conversational Systems in Indian Languages (IECSIL) which has been organized by FIRE 2018. Motivated by the need of Information Extractor, corpora has been developed to perform the Named Entity Recognition (Task A) and Relation Extraction (Task B) for five Indian languages (Hindi, Tamil, Malayalam, Telugu and Kannada). Task A is to identify and classify the named entities to one of the many classes and Task B is to extract the relation among the entities present in the sentences. Altogether, nearly 100 submission of 10 different teams were evaluated. In this paper, we have given an overview of the approaches and also discussed the results that the participated teams have attained.

Keywords: Information Extractor · Named Entity Recognition · Relation Extraction · IECSIL.

1 Introduction

Applications of conversational systems and social media platforms have seen increased adoption by Indian language users on account of local language enabled keyboards and smart phones [3]. In recent times, e-tailing, digital classifieds, digital payments and on-line government services have also started to enable Indian language content on their platforms. This growth momentum is likely to continue with the Indian language Internet user base growing at a CAGR of 18% to reach 536 million by 2021 compared to English Internet user base growing at 3% to reach 199 million. Their study shows that by 2021, almost all domains would be benefited with the support of their own local language and

there would a drastic increase in the amount of data that gets generated when compared to the present case. More research works and state-of-art findings are likely to happen in near future. Researchers and Start-ups have already started following up the need for language support in frequently used applications which would in turn benefit most of the crowd in India.

Understanding the above scenarios, Arnekt in collaboration with FIRE has come up with a track Arnekt-IECSIL - Information Extractor for Conversational Systems in Indian Languages (IECSIL). FIRE started of with the aim of building a South Asian counterpart for TREC, CLEF and NTCIR. FIRE has since evolved continuously to meet the new challenges in multilingual information access.[4]. Arnekt aims to power the world's smartest business solutions by providing state-of-the-art AI based Cognitive Intelligence as a Service (CIaaS)[5]. IECSIL basically involves five Indian languages (Hindi, Kannada, Malayalam, Tamil and Telugu) to start with and is likely to be further extended to cover the major languages spoken in India (near future).

Resources for developing this prototype was collected using an automated and language independent framework which has been developed by Arnekt, that creates corpus for Named Entity Recognition (NER) and Relation Extraction (RE) (tasks in IECSIL) from DBpedia. Corpora contains tags of Named Entities and Relations for five Indian languages (Kannada (kn), Malayalam (ml), Hindi (hi), Tamil (ta) and Telugu (te)) which are not just restricted towards creating a single application. An elaborated portion on steps taken for data creation and its statistics could be seen below in the coming sections.

Motivated by the need of Information Extractor described above, we have the following two tasks:

Task A : Named Entity Recognition (NER)

Corpora for five Indian languages (Hindi, Tamil, Malayalam, Telugu and Kannada) has been provided. Task A is to identify and classify the named entities to one of the many classes [2].

NER Corpus Creation: The abstract and info-box property files from DBpedia are the resources for corpus creation. In preprocessing stage, info-box properties are extracted as a meta tags and the long abstract files are cleaned to remove the texts in foreign language, URL links, and other special symbols. The meta tags which are in non-English language has been translated into English through Google Translator. The meta tags that occurs more than 100 times across all the languages has been considered to create the final entity and its corresponding text pairs. With this entity-text pair, the text in the cleaned abstract file has been tagged. There are totally nine tags (Date, Event, Location, Name, Number, Occupation, Organization, Other and Things) which are considered for the NER corpus creation.

[4] http://fire.irsi.res.in/fire/2018/home
[5] https://arnekt.com/

Creation of meta tag to the entity list is the only manual processing involved in this framework and it takes very less time compared to the general manual annotation process. This corpus has been made available on-line[6] to the research community through the Information Extractor for Conversational Systems for Indian Languages (IECSIL)[7]. The detailed NER corpus statistics has been given in Table 1:NER Corpus Statistics.

Table 1. NER Corpus Statistics

Info	**Languages**				
	hi	**kn**	**ml**	**ta**	**te**
date	4290	1968	2606	24556	3999
event	4968	916	1432	8439	1230
location	278396	17484	49705	225229	159840
name	149300	25576	101914	202120	103256
number	63289	6519	51122	130581	47727
occupation	26418	5136	13462	27398	14188
organization	20831	1237	8078	16601	4156
other	1903703	439238	1167211	1844116	959260
things	6804	389	3435	10244	1855

Task B : Relation Extraction (RE)

Continuation to Task A, corpora without named entities for five Indian languages (Hindi, Tamil, Malayalam, Telugu and Kannada) has been provided. Task B is to extract the relation among-st the entities present in the sentences [1] .

Relation Extraction Corpus Creation: Similar to NER, here also relation tags are annotated through semi-automated methodology. Initially sentence which has minimum NER tags count two has been taken and POS tagging is applied on it. The tagger from the [5] [4] and [6] are used to create the POS tagged corpus for all five languages. The POS tags from these tools are mapped to the commonly occurring 12 Penn Treebank POS tags, which are good enough to use it in the further application. Based on the POS pattern between the entities, each sentence is assigned to a relation [1]. The relation tagged corpus statistics is given in Table 2.

2 Evaluation

For evaluation, the classic Accuracy measure has been taken into consideration. It could simply be briefed as a predictive model that reflects the proportionate

[6] https://github.com/BarathiGanesh-HB/ARNEKT-IECSIL
[7] http://iecsil.arnekt.com

Table 2. Relation Extraction Corpus Statistics

Info	Languages				
	hi	**kn**	**ml**	**te**	**ta**
action_1	15517	0	0	0	1974
action_2	740	277	340	2150	4512
action_3	9	321	2260	1306	2661
action_neg	199	0	0	0	78
action_per	3	9	1056	23	222
action_so	0	0	248	0	0
action_quant	70	25	152	13	14
information_1	29264	2918	13854	8550	38569
information_2	34	172	1990	15078	815
information_3	469	807	4068	3539	3681
information_4	5388	342	337	1113	1544
information_cc	80	102	0	268	142
information_closed	2063	3	148	1030	786
information_neg	6	4	0	0	135
information_per	443	869	1225	1641	3125
information_quant	907	414	931	1650	4577
information_so	0	0	0	115	969
Other	1583	374	1678	563	1029

number of times that the model is correct when applied to data. Evaluation has been computed in two stages,

Pre-Evaluation

Team participating in the shared tasks were encouraged to test their modules in real time[8]. They could feel free in submitting as many submissions as they prefer. The leader board is evaluated with approximately 20% of the data (Test-1 corpora). Test-1 corpora statistics are given in Table 3 and 4.

Final-Evaluation

The final ranking is based on another 20% (Test-2 corpora) of the data. Unlike the Pre-Evaluation, here the participants are requested to submit their models or code or submission file to task organizers. Test-2 corpora statistics are given in Table 3 and 4.

For each sub-task and language, submissions are evaluated by calculating the accuracy with the corresponding Gold labels. The accuracy scores across all the five languages will be averaged to determine the final ranking for both the sub-tasks.

[8] https://iecsil.arnekt.com/#!/participate

$$Acc = \frac{\#\ terms\ correctly\ assigned\ to\ entity}{total\ \#\ terms} \quad (1)$$

Table 3. Task A Corpus Separation : NER

TASK - A			
Language	**Train**	**pre-Eval**	**final-Eval**
hi	1548570	519115	517876
kn	318356	107325	107010
ml	903521	301860	302232
te	840908	280533	279443
ta	1626260	542225	544183

Table 4. Task B Corpus Separation : Relation Extraction

TASK - B			
Language	**Train**	**pre-Eval**	**final-Eval**
hi	56775	18925	18926
kn	6637	2213	2213
ml	28287	9429	9429
te	37039	12347	12347
ta	64833	21611	21612

3 Participants

A server similar to Kaggle/Coda Lab was hosted[9] to check the developed system in real time, where participants submitted their test results for pre-evaluation corpora. Five days before the final deadline Test 2 corpora for final evaluation has been released. Participants were allowed to make at most 3 submissions against the Test 2 corpora. The final ranking was then computed based on the participants system performance on Test 2 corpora. The results are described in Table 5, 6, 7 and 8.

The **CUSAT_TEAM** have made use of deep learning in extracting the relation between entities. They have used Convolutional Neural Network (CNN), which has been modelled to address processing in sentence level for Malayalam language. Due to the absence of pre-trained word embedding for other languages

[9] https://iecsil.arnekt.com/#!/participate

Table 5. Pre-Evaluation Task A

Team	hi	kn	ml	ta	te	Average
idrbt-team-a	97.82	97.04	97.46	97.41	97.54	97.45
CUSAT_TEAM	97.67	97.03	97.44	97.36	97.72	97.44
rohitkodali	98.07	96.86	97.26	96.98	97.54	97.34
khushleen	96.84	96.38	96.64	96.15	96.63	96.53
thenmozhi	96.73	95.63	95.87	95.55	96.77	96.11
hariharan-v	96.49	95.06	95.9	96.03	95.97	95.89
hilt	94.44	92.94	92.92	92.48	92.42	93.04
am905771	94.4	90.09	89.97	91.23	90.2	91.18
raiden11	91.52	92.14	90.27	87.72	90.02	90.33

Table 6. Pre-Evaluation Task B

Team	hi	kn	ml	ta	te	Average
thenmozhi	93.25	51.20	81.89	85.91	84.29	79.30
idrbt-team-a	80.98	57.98	59.43	78.43	76.35	70.63
raiden11	51.70	44.42	48.61	59.71	40.57	49.00
hilt	51.70	44.42	48.61	59.71	21.97	45.28
CUSAT_TEAM	51.70	0	78.45	0	0	26.03
am905771	63.74	0	0	0	0	12.75

like Hindi, Kannada, Tamil and Telugu that fits in to their machine memory, they have restricted their Relation extraction model development with Malayalam language for which they have their own corpus to simulate word vectors. The same team have used a statistical model in finding the entities from a given sentence. CRF based sequence labelling model with features that are specific to Indian languages has been utilized in tagging the words with entities provided [7], [8].

SSN_NLP have used Neural Machine Translation architecture to identify and classify named entities for all the five Indian languages that are in focus. The deep neural network was built using multi-layer Recurrent Neural Network (RNN) and Long Short Term Memory (LSTM). About four different models were developed for each of the languages. It was found that bi-directional LSTM with attention having eight layers of depth worked well for all languages other than Malayalam [9].

SSN_NLP have made use of the deep learning approach that they have utilized for Named Entity Recognition (NER) for Relation Extraction as well. While two models use the deep learning framework that use SeqtoSeq model, three others were developed using statistical Machine Learning approach [10].

HiLT have used two-layer Convolutional Neural Network (CNN) for character level (word-matrix) and word level encoding (sentence-matrix), along with a Bidirectional Long Short Term Memory (Bi-LSTM) as a tag decoder for Named Entity Recognition. This non-linear model has been developed as a language independent framework with the aim of extending it to other Indian languages

other than the five languages in focus. It is an added advantage that their model does not seem to be biased for a particular language [11].

IIT(BHU) generated vector representation of words and their corresponding tags, that were fed to the Bidirectional Long Short Term Memory (Bi-LSTM) for identification and categorization of entities in the text. Word representation has been done for all possible words in the corpus and a set of unique words were represented using one-hot encoding. The BiLSTM layer here learns the contextual relationship between words from past and future context. This team has come up with a language independent framework for Named Entity Recognition (NER) and has proven the same for the five languages provided [12].

Khushleen has made use of character level information in order to include word representation for rare words or out of vocabulary words from the given corpora. The team has performed word embedding using fastText without changing the parameters for each language for building a unified model. This is then fed to a two-layer Bidirectional Long Short Term Memory (BiLSTM) for training and prediction of entities for words in sentences [13].

Table 7. Final Evaluation Task A

Team	Run	ml	kn	hi	ta	te
hilt	2	92.1	93.17	94.35	91.79	92.47
raiden11	1	89.6	92.33	91.19	87.26	89.19
SSN_NLP	3	95.05	94.21	95.95	94.66	95.4
hilt	2	92.12	93.17	94.28	91.79	92.47
am905771	2	88.89	89.85	94.47	90.4	90.04
idrbt-team-a	1	96.58	96.79	97.82	96.18	97.68
SSN_NLP	2	95.28	95.76	96.51	94.9	96.81
khushleen	1	96.18	96.45	96.85	95.83	96.78
CUSAT_TEAM	1	96.86	97.09	97.65	96.85	97.69
hariharanv	1	95.63	95.79	96.67	NA	96.39
rohitkodali	1	NA	96.85	98.06	NA	97.53
am905771	3	89.13	89.88	94.92	90.47	90.32
SSN_NLP	1	95.28	95.8	96.68	94.91	96.81
am905771	1	89.04	89.53	94.45	90.46	90.04

Semantic relation among-st words were captured using word embedding as done in Khushleen work using fastText by the **Raiden11** team. As a next step they have experimented this work using linear models like Naive Bayes and Support Vector Machine. Apart from this they were able to prove that a simple Artificial Neural Network (ANN) model worked better than the former linear classifiers, as it could capture the composite relation between words [13].

idrbt-team-a used a two stage LSTM based network with character based emebeddings, word2vec embeddings and sequence based bi-LSTM embeddings together to carry all the requisite features necessary for the NER prediction problem [14].

In Relation Extraction, the team **idrbt-team-a** used features like POS tags, NER tags along with the words in input text sentence to classify the given input into one of the predefined relationship class. By performing the initial experiment with other statistical classifiers, Logistics Regression is chosen as the classifier [15].

By using word embedding from fastText as a representation method, team **raiden11** have experimented the linear models like Naive Bayes and SVM, and also a simple Neural Network to develop the NER system. The best results are achieved for neural network for all languages combined [16].

Table 8. Final Evaluation Task B

Team	Run	ml	kn	hi	te	ta
CUSAT_TEAM	1	77.77	NA	NA	NA	NA
hilt	2	48.05	44.01	51.5	22.87	60.11
idrbt-team-a	1	57.86	57.34	79.21	76.14	78.44
raiden11	1	48.05	44.01	51.5	40.49	60.11
SSN_NLP	1	81.99	51.87	92.99	84.11	86.26
hilt	1	48.05	44.01	51.5	22.87	60.11
SSN_NLP	3	51.8	49.43	69.04	68.17	67.12
SSN_NLP	2	75.25	45.14	91.71	85.78	82.19

Participants were mostly used deep learning based algorithms for both the Relation Extraction and Named Entity Recognition tasks. CNN, Bi-LSTM and CNN with Bi-LSTM are commonly used architectures. Participants yields 90 $\pm$ 5 % as the accuracy for NER task. Even though the accuracy is high, it has to be noted that the accuracy obtained by selecting all entity as the class **"other"** is 80 $\pm$ 5 %. This can be observed by measuring the performance of the team through f1 score.

Unlike NER, participated systems could not able to attain the best results. The above points shows the need of research in Indian Language based NER and Relation Extraction systems. The detailed results including the precision, recall and f1 score for target class and language is made publicly available [10].

4 Conclusion

Arnekt in collaboration with FIRE has come up with its first track on Information Extraction for Conversational Systems in Indian Languages (IECSIL), which has utilized five Indian languages (Hindi, Kannada, Malayalam, Tamil and Telugu) for identifying the entities (Task A : Named Entity Recognition) and also extracting relation from the same (Task B : Relation Extraction). IECSIL has developed its own corpora for both the tasks. While this corpus is not

[10] https://github.com/BarathiGanesh-HB/ARNEKT-IECSIL/blob/master/IECSIL-2018-Final-Evaluation-Results.xlsx

restricted for a single application, it has been made available on-line[11] to the research community through the Information Extractor for Conversational Systems for Indian Languages (IECSIL)[12]. The teams who have participated have come up with feasible solutions and most of them have utilized Deep learning methods to build their models. With the increase in need of Indian language usage, we are likely to extend the number of Indian languages used in the near future.

5 ACKNOWLEDGEMENTS

Arnekt thanks all the participants for showing their interest towards IECSIL. We would also like to show are gratitude to the FIRE 2018 organizers for their endless efforts and support.

References

1. Bhatt B, Bhattacharyya P. Domain specific ontology extractor for indian languages. InProceedings of the 10th Workshop on Asian Language Resources 2012 (pp. 75-84).
2. Nayan A, Rao BR, Singh P, Sanyal S, Sanyal R. Named entity recognition for Indian languages. InProceedings of the IJCNLP-08 Workshop on Named Entity Recognition for South and South East Asian Languages 2008.
3. Zamora J. Rise of the chatbots: Finding a place for artificial intelligence in India and US. InProceedings of the 22nd International Conference on Intelligent User Interfaces Companion 2017 Mar 7 (pp. 109-112). ACM.
4. Tamil Shallow Parser, International Institute of Information Technology, Hyderabad, https://ltrc.iiit.ac.in/showfile.php?filename=downloads/shallow_parser.php.
5. Reddy S, Sharoff S. Cross language POS taggers (and other tools) for Indian languages: An experiment with Kannada using Telugu resources. InProceedings of the Fifth International Workshop On Cross Lingual Information Access 2011 (pp. 11-19).
6. Devadath VV, Sharma DM. Significance of an accurate sandhi-splitter in shallow parsing of dravidian languages. InProceedings of the ACL 2016 Student Research Workshop 2016 (pp. 37-42).
7. Ajees A P and Sumam Mary Idicula, CUSAT_TEAMIECSIL-FIRE-2018:A Named Entity Recognition System for Indian Languages, FIRE Working Notes, 2018.
8. Ajees A P and Sumam Mary Idicula, CUSAT_TEAMIECSIL-FIRE-2018: A Relation Extraction System for Indian Languages, FIRE Working Notes, 2018.
9. D. Thenmozhi, B. Senthil Kumar, and Chandrabose Aravindan, SSN_NLPIECSIL-FIRE-2018: Deep Learning Approach to Named Entity Recognition for Conversational Systems in Indian Languages, FIRE Working Notes, 2018.
10. D. Thenmozhi, B. Senthil Kumar, and Chandrabose Aravindan, SSN_NLPIECSIL-FIRE-2018: Deep Learning Approach to Relation Extraction for Conversational Systems in Indian Languages, FIRE Working Notes, 2018.
11. Sagar, Srinivas P Y K L, Rusheel Koushik Gollakota, and Amitava Das, HiLTIECSIL-FIRE-2018, FIRE Working Notes, 2018.

[11] https://github.com/BarathiGanesh-HB/ARNEKT-IECSIL
[12] http://iecsil.arnekt.com

12. Akanksha Mishra, Rajesh Kumar Mundotiya, and Sukomal Pal, IIT(BHU)IECSIL-FIRE2018: Language Independent Automatic Framework for Entity Extraction in Indian Languages, FIRE Working Notes, 2018.
13. Khushleen Kaur, KhushleenIECSIL-FIRE-2018:Indic Language Named Entity Recognition Using Bidirectional LSTMs with Subword Information, FIRE Working Notes, 2018.
14. S. Nagesh Bhattu, N. Satya Krishna , and D. V. L. N. Somayajulu, idrbt-team-aIECSIL-FIRE-2018 Named Entity Recognition of Indian languages using Bi-LSTM, FIRE Working Notes, 2018.
15. N. Satya Krishna , S. Nagesh Bhattu , and D. V. L. N. Somayajulu, idrbt-team-aIECSIL-FIRE-2018 : Relation Categorisation for Social Media News Text , FIRE Working Notes, 2018.
16. Ayush Gupta, Meghna Ayyar, Ashutosh Kumar Singh, and Rajiv Ratn Shah, raiden11IECSIL-FIRE-2018 : Named Entity Recognition For Indian Languages, FIRE Working Notes, 2018.