ICON 2020

17th International Conference on Natural Language Processing

Proceedings of the System Demonstration

December 18 - 21, 2020
Indian Institute of Technology Patna, India

Introduction

Welcome to the proceedings of the system demonstration session. This volume contains the papers of the system demonstrations presented at the 17th International Conference on Natural Language Processing, held virtually organized by AI-NLP-ML Group, IIT, Patna, on December 18 2020.

The system demonstrations program offers demonstrations of complete working NLP Systems. The system demonstration chair and the members of the program committee received 25 submissions, 18 of which were selected for inclusion in the program after reviewing by two members of the program committee.

I would like to thank the members of the program committee for their excellent job in reviewing the submissions and providing their support for the final decision.

Chairs

Vishal Goyal (Punjabi University, Patiala), Asif Ekbal (IIT Patna)

Program Committee:

Gurpreet Singh Lehal, Punjabi University Patiala, India
Sanjay Dwivedi, Babasaheb Bhimrao Ambedkar University, Lucknow, U.P, India
Amba Kulkarni, University of Hyderabad, Hyderabad, India.
Rajeev R R, ICFOSS, Trivandrumpuram, India.
Neeraj Mogla, Facebook, USA.
Sanjeev Gupta, Google, Bangalore.

Additional Reviewers:

Aditi Sharan, Jawaharlal Nehru University, New Delhi, India
Sanjeev Sharma, DAV University, Jalandhar, India.
Gurpreet Singh Josan, Punjabi University Patiala, India.
Parminder Singh, Guru Nanak Dev Engineering College, Ludhiana, India
Basant Aggarwal, IIIT, Kota.
Manish Shrivastava, IIIT, Hyderabad.

Table of Contents

System Demonstration Program

Monday, December 21, 2020

+ 10:00 - 10:15 **Opening Remarks by Dr. Vishal Goyal**

+ 10:15 - 12:55 **Session I**

NLP Tools for Khasi, a low resource language
Medari Janai Tham

The Hindi to Dogri Machine Translation System
Preeti Dubey

Plagiarism detection tool for Indian Language documents with Special Focus on Punjabi and Hindi Language
Vishal Goyal, Rajeev Puri, Jitesh Pubreja and Jaswinder Singh

Railway Stations Announcement System for Deaf
RAKESH KUMAR, Vishal Goyal and Lalit Goyal

A Chatbot in Malayalam using Hybrid Approach
PRAVEEN PRASANNAN, Stephy Joseph and Rajeev R R

Airport Announcement System for Deaf
RAKESH KUMAR, Vishal Goyal and Lalit Goyal

Automatic Translation of Complex English Sentences to Indian Sign Language Synthetic Video Animations
Deepali, Vishal Goyal and Lalit Goyal

Sentiment Analysis of English-Punjabi Code-Mixed Social Media Content
Mukhtiar Singh, Vishal Goyal, Sahil Raj

Language Identification and Normalization of Code Mixed English and Punjabi Text
Neetika Bansal, Vishal Goyal and Simpel Rani

Software to Extract Parallel Data from English-Punjabi Comparable Corpora
Manpreet Singh Lehal, Ajit Kumar, Vishal Goyal

Development of Hybrid Algorithm for Automatic Extraction of Multiword Expressions from Monolingual and Parallel Corpus of English and Punjabi
Kapil Dev Goyal and Vishal Goyal

+ 14:00 - 15:40 **Session 2**

Design and Implementation of Anaphora Resolution in Punjabi Language
Kawaljit Kaur, Vishal Goyal and Kamlesh Dutta

Punjabi to English Bidirectional NMT System
Kamal Deep, Ajit Kumar and Vishal Goyal

Urdu To Punjabi Machine Translation System
Umrinder Pal Singh, Vishal Goyal and Gurpreet Lehal

Punjabi to Urdu Machine Translation System
Nitin Bansal and Ajit Kumar

Opinion Mining System for Processing Hindi Text for Home Remedies Domain
ARPANA PRASAD, Neeraj Sharma and Shubhangi Sharma

Demonstration of a Literature Based Discovery System based on Ontologies, Semantic Filters and Word Embeddings for the Raynaud Disease-Fish Oil Rediscovery
Toby Reed and Vassilis Cutsuridis

A Sanskrit to Hindi Language Machine Translator using Rule Based Method
Prateek Agrawal and Vishu Madaan

+ 15:40 - 16:00 **Vote of Thanks by Dr. Asif Ekbal**

Demonstration of a Literature Based Discovery System based on Ontologies, Semantic Filters and Word Embeddings for the Raynaud Disease-Fish Oil Rediscovery

Toby Reed[1,2], Vassilis Cutsuridis[1]
[1] School of Computer Science, University of Lincoln, Lincoln, LN6 7TS, UK.
[2] Streets Heaver Healthcare Computing, The Point, Weaver Rd, Lincoln LN6 3QN, UK
toby.reed@streets-heaver.com, vcutsuridis@lincoln.ac.uk

Abstract

A novel literature-based discovery system based on UMLS Ontologies, Semantic Filters, Statistics, and Word Embeddings was developed and validated against the well-established Raynaud's disease – Fish Oil discovery by mining different size and specificity corpora of Pubmed titles and abstracts. Results show an 'inverse effect' between open versus closed discovery search modes. In *open* discovery, a more general and bigger corpus (Vascular disease or Perivascular disease) produces better results than a more specific and smaller in size corpus (Raynaud disease), whereas in *closed* discovery, the exact opposite is true.

1 Introduction

In the current COVID-19 era there is widespread demand from the pharmaceutical and healthcare industries for more work to be done in the field of reusing compounds in diseases as a method to escape some of the most expensive and time-consuming processes in drug discovery. After the most famous Sildenafil (Viagra) being repurposed from cardiovascular disease to erectile dysfunction, the use of drug repositioning has been shown to have the potential to be beneficial not only to the healthcare facilities and pharmaceutical companies, but also to the everyday consumer particularly if the process to finding and developing cures becomes cheaper, then the actual to consumer cost of treatment will likely decrease (Reed, 2020). One method for drug repositioning is through Literature Based Discovery (LBD), a powerful text mining approach that harnesses already available scientific knowledge to build bridges between seemingly unrelated islands of knowledge, such as the association of an existing drug to a novel medical condition (Reed, 2020). LBD is classified into two types: *open* and *closed* discovery. In closed discovery (also known as *hypothesis testing*), the user specifies a pair of topics (A and C) and the objective is to find any unknown, but meaningful connections (the *intermediate* (B) terms) between them. In open discovery (also known as *hypothesis generation*), the user specifies a topic of interest (C) (e.g. a disease or a drug) and the system finds a set of *intermediate* (B) terms directly related to the starting topic of interest. For each of these intermediate terms, the system reiterates the same mechanism to generate a set of *final* (A) terms.

2 Materials and Methods

A novel LBD system based on Word Embeddings, Statistics, Semantic Filters, and UMLS ontologies was developed to rediscover the Raynaud disease-Fish Oil connection (Swanson, 1986) by mining Pubmed titles and abstracts. Our system's pipeline and corpora mined to discover the Raynaud disease – Fish oil connection (Swanson, 1986) can briefly be described as follows:

1. Corpora: Different size and specificity corpora of Pubmed titles/abstracts were retrieved for each discovery type (open vs closed). *Open discovery corpora*: (i) Vascular disease, (2) Peripheral vascular disease (PVD), and (3) Raynaud disease. *Closed*

Proceedings of the 17th International Conference on Natural Language Processing: System Demonstrations, pages 1–3
Patna, India, December 18 - 21, 2020.

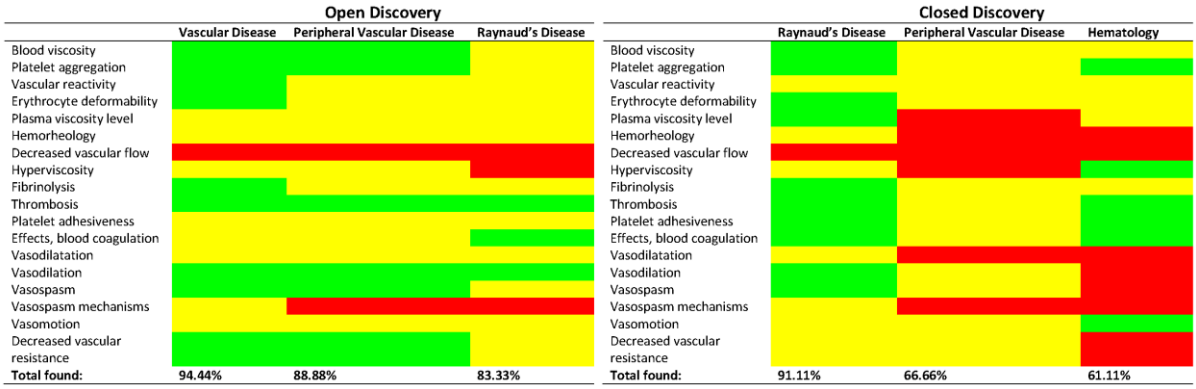

Figure 1: Discovered B-terms from the three open-discovery (*Left*) and closed-discovery (*Right*) corpora. In 'green' are the correctly rediscovered concepts, in 'yellow' the semantically similar discovered concepts, and in 'red' the concepts our system ought to have discovered but failed to do so.

discovery corpora: (i) Hematology, (ii) PVD, and (3) Raynaud disease. Raynaud disease is a specific type of vascular disease, but not a type of PVD, which is a sub-type of vascular disease. Raynaud disease is a sub-type of vascular disease, which involves blood, but not per se a sub-type of a hematological disease and neither is PVD.

2. Pre-processing: Each retrieved title/abstract of a scientific article was normalized to remove word variations due to capitalization. Any words with the less than three characters was also removed from further processing. All remaining words were then passed through a Natural Language Toolkit parser to generate bigrams/trigrams of each unigram based on a minimum occurrence count value.
3. A Skip-Gram Word2Vec model (Mikolov et al., 2013) was employed with some initial parameter values to generate word vectors for all words and phrases in each corpus.
4. We scanned through all generated word vectors to discover variations of the "raynaud" C-concept (e.g. Raynaud's disease, Raynaud syndrome, primary Raynaud, etc).
5. We utilised a grid search on the architecture, dimensionality, epoch, learning rate, down-sampling, context window and minimum word count parameters to find the model with the optimum performance in each corpus used.
6. Using the optimally derived Word2Vec model, we repeated STEP 4 to estimate cosine similarity of all B- or A-terms in the corpus with Raynaud variation terms from STEP 3.
7. Placed the most semantically similar terms with the closest cosine similarity, from STEP 5 into a list.
8. Mapped every term from the list via MetaMap (Aronson and Lang, 2010) to UMLS ontologies (Bodenreider, 2004). Using a semantic filter we excluded from further analysis all mapped terms which were not semantically related to the semantic types in the filter.
9. These results were then compared to previously found terms to see if our system provided acceptable results.

3 Results and Discussion

In Figure 1 results from both discovery modes. show an 'inverse effect'. In closed discovery a more specific, but smaller in size corpus (Raynaud disease) produced better results than a more general and bigger in size corpus (PVD or Hematology). On the contrary, in open discovery, a more general and bigger corpus (Vascular disease or PVD) produced better results than a more specific and smaller in size corpus (Raynaud disease). This result indicates to detect hidden relations between domain specific terms in just one-step (A-B-C), which otherwise is a multi-step process (A-B, B-C, A-C), is preferable to have more general amounts of data of the targeted domain problem to extend much further from it than to have specific data uncovering only one of these relationships (A-B or B-C). In contrast, to automatically detect words that are domain specific, is preferable to have a corpus that correctly represents the use of these specific concepts than to have more general amounts of data that encapsulate the targeted domain problem, but extend much further from it.

References

Toby S Reed. 2020. *Use of Word Embeddings in a Literature-Based Discovery System*. Master by Research Thesis, University of Lincoln, Lincoln, UK

Don R Swanson. 1986. Fish oil, Raynaud's syndrome, and undiscovered public knowledge. *Perspectives in biology and medicine,* 31(4): 7-18

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, Jeffrey Dean. 2013. Efficient estimation of words and phrases and their compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems (NIPS) 2013*, pages 3111-311, Lake Tahoe, Nevada.

Alan A Aronson and Francois-Michel Lang. 2010. An overview of Metamap: historical perspective and recent advances. *Journal of the American Medical Informatics Association,* 17: 229-236.

Olivier Bodenreider. 2004. The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nuclei Acids Research,* 32: D267-D270.

Development of Hybrid Algorithm for Automatic Extraction of Multiword Expressions from Monolingual and Parallel Corpus of English and Punjabi

Kapil Dev Goyal, Vishal Goyal
Department of Computer Sciecne,
Punjabi University Patiala
{kapildevgoyal,vishal.pup}@gmail.com

Abstract

Identification and extraction of Multiword Expressions (MWEs) is very hard and challenging task in various Natural Language processing applications like Information Retrieval (IR), Information Extraction (IE), Question-Answering systems, Speech Recognition and Synthesis, Text Summarization and Machine Translation (MT). Multiword Expressions are two or more consecutive words but treated as a single word and actual meaning this expression cannot be extracted from meaning of individual word. If any systems recognized this expression as separate words, then results of system will be incorrect. Therefore it is mandatory to identify these expressions to improve the result of the system. In this report, our main focus is to develop an automated tool to extract Multiword Expressions from monolingual and parallel corpus of English and Punjabi. In this tool, Rule based approach, Linguistic approach, statistical approach, and many more approaches were used to identify and extract MWEs from monolingual and parallel corpus of English and Punjabi and achieved more than 90% f-score value in some types of MWEs.

1 Introduction

In this tool, ruled based, linguistic and statistical approaches are used to extract MWEs. Apart from these approaches, Part of Speech tagger tool, Named Entities Recognizer tool, Stemmer, Giza++, etc tools are used. Most of MWEs are generic in nature, it means based on rules and linguistic approach than statistical approach. Mostly ruled based approach is used in Replicated words, which are strong candidate of MWEs in Punjabi Language. The results of linguistic approach in English are better than results of Punjabi, because of the poor performance of Punjabi Part of Speech tagger tool and Punjabi Stemmer. Similarly results of monolingual corpus are better than parallel corpus, because of the lack of proper Punjabi-English dictionary and poor performance of giza++ tool. Therefore results of MWEs are directly depending upon performance of above mentioned tools. As we earlier discussed, most of the NLP applications are highly affected by MWEs. This automatic MWEs tool will help the performance of many NLP applications like Information Retrieval (IR), Information Extraction (IE), Question-Answering systems, Speech Recognition and Synthesis, Text Summarization and Machine Translation (MT). Non-compositional, non-modifiable and non-substitutable are basic features of MWEs. Non Compositional means that meaning of MWEs cannot be predicted from meaning its parts. Non Modifiable means that Multiword Expressions are frozen and they cannot be changed in any way. Non Substitutable means that any parts of Multiword Expression cannot be substituted by one of its synonym without affecting the meaning of an expression.

1.1. Features of MWEs

(Manning & Schutze, 1999) described that non-compositional, non-modifiable and non-substitutable are basic features of MWE.

(1) Non-compositional: It means that MWE cannot be completely translated from the meaning of its parts.

E.g. ਅੱਖਾਂ ਦਾ ਤਾਰਾ) Punjabi)

Transliteration: “Akhān dā tārā”
Gloss: Star of Eyes
Translation: Lovely

E.g. लोहे के चने चबाना (Hindi)

Transliteration: “Lōhē kē chanē chabānā”

Proceedings of the 17th International Conference on Natural Language Processing: System Demonstrations, pages 4–6
Patna, India, December 18 - 21, 2020.

Gloss: To chew iron gram
Translation: Difficult task

In above examples, actual translations cannot be predicted from their parts, which are completely different from its basic meaning.

(2) Non-modifiable:
Many Multiword Expressions are frozen and they cannot be changed in any way. These types of expressions cannot be modified by grammatical transformations (like by changing Number/ Gender/ Tense, addition of adjective etc).
Eg. In ਰੋਜੀ ਰੋਟੀ) Rōjī rōṭī) cannot be changed in number as ਰੋਜੀ ਰੋਟੀਆਂ) Rōjī rōṭī'ān)

(3) Non-Substitutable:
Any word of Multiword Expression cannot be substituted by one of its synonym without affecting the meaning of an expression.
E.g. ਰੋਜੀ ਰੋਟੀ) Rōjī rōṭī) cannot be written as ਰੋਜੀ ਖਾਣਾ) Rōjī khāṇā) or ਰੋਜ ਰੋਟੀ) Rōj rōṭī)

2 Review of literature

The concept of Multiword Expression is given by (Baldwin & Kim, 2010; Sag et al., 2002) has covered all types of MWEs. Identification and extraction of MWEs is not very old field, but still many of the researchers are working in this field. There has been very limited work done reported on monolingual Punjabi MWEs and extraction of parallel MWEs from English-Punjabi parallel corpus. Most of researcher used statistical method or association measures tools (Evert & Krenn, 2005), linguistic based approaches (Goldman et al., 2001; Vintar & Fišer, 2008), ruled based approaches (M Nandi, 2013), hybrid approaches (Boulaknadel et al., 2008; Jianyong et al., 2009) for extracting MWEs expression from monolingual and parallel corpus. In Indian language, (Rácová, 2013; Singh et al., 2016; Sinha, 2009, 2011) classified reduplicate MWEs which are strong candidates of MWEs.
3 Methodology
This report presents an automatic tool which extracts following Multiword Expressions from English and Punjabi Language.

3.1. Punjabi Multiword Expressions:

3.1.1. Replicated Multiword Expression
Replicated MWEs
Combination with antonyms/gender
Combination with hyponyms
'Walaa' Morpheme
3.1.2. Extracted using Statistical Methods
3.1 3. Extracted using Linguistic Methods
Name Entities
Compound Noun
Conjunct Verbs
Compound Verbs
3.1.4. Extracted using Hybrid Approach

3.2. English Multiword Expressions:

3.2.1. Extracted using Statistical Methods
3.2.2. Extracted using Linguistic Methods
Name Entities
Compound Noun
Conjunct Verbs
Compound Verbs
3.2.3. Extracted using Hybrid Approach

3.3. Punjabi-English Parallel Multiword Expressions:

Replicated Multiword Expression
Extracted using Statistical Methods
Extracted using Linguistic Methods
Extracted using Hybrid Approach

4 Results

MWEs Type	Accuracy	Precision	Recall	F-Score
Rule Based Approach	99.88	94.73	81.81	87.80
Using Linguistic Methods	99.67	33.33	75	46.15
Using Statistical Methods	80.03	52.35	29.29	37.56

In these results, accuracy scores are more than 70%, but precision, recall and F-score values are varied from 30% to 97%. Results of replicated words using rule based approach are relatively better than linguistic approach and statistical approach. Statistical tools measure the association between words, therefore results for statistical methods are relatively less than all above types.

References

Baldwin, T., & Kim, S. N. (2010). Multiword expressions. *Handbook of Natural Language Processing, Second Edition*, 267–292.

Boulaknadel, S., Daille, B., & Aboutajdine, D. (2008). A multi-word term extraction program for Arabic language. *Proceedings of the 6th International Conference on Language Resources and Evaluation, LREC 2008*.

Evert, S., & Krenn, B. (2005). Using small random samples for the manual evaluation of statistical association measures. *Computer Speech and Language*, *19*(4). https://doi.org/10.1016/j.csl.2005.02.005

Goldman, J.-P., Nerima, L., & Wehrli, E. (2001). Collocation extraction using a syntactic parser. *Proceedings of the ACL Workshop on Collocations*, 61–66.

Jianyong, D., Lijing, T., Feng, G., & Mei, Z. (2009). A hybrid approach to improve bilingual multiword expression extraction. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. https://doi.org/10.1007/978-3-642-01307-2_51

M Nandi, R. R. (2013). Rule-based Extraction of Multi-Word Expressions for Elementary Sanskrit. *International Journal of Advanced Research in Computer Science and Software Engineering*.

Manning, C., & Schutze, H. (1999). *Foundations of statistical natural language processing*. https://books.google.com/books?hl=en&lr=&id=3qnuDwAAQBAJ&oi=fnd&pg=PT12&dq=Foundations+of+statistical+natural+language+processing.+MIT+press,+1999.+manning&ots=ysF-mZAwM_&sig=GplFqEroiO9dy1ZGYhebtAhhEdk

Rácová, A. (2013). Reduplication of verbal forms in Bengali. *Asian and African Studies*.

Sag, I. A., Baldwin, T., Bond, F., Copestake, A., & Flickinger, D. (2002). Multiword expressions: A pain in the neck for NLP. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *2276*, 1–15. https://doi.org/10.1007/3-540-45715-1_1

Singh, R., Ojha, A. K., & Jha, G. N. (2016). *Classification and Identification of Reduplicated Multi-Word Expressions in Hindi*. *May*.

Sinha, R. M. K. (2009). *Mining complex predicates in Hindi using a parallel Hindi-English corpus*. *August*, 40. https://doi.org/10.3115/1698239.1698247

Sinha, R. M. K. (2011). Stepwise mining of multi-word expressions in Hindi. *Workshop on Multiword Expressions: From Parsing and Gerenraion to Real World(MWE 2011)*, *June*, 110–115. http://dl.acm.org/citation.cfm?id=2021121.2021143

Vintar, Š., & Fišer, D. (2008). Harvesting multi-word expressions from parallel corpora. *Proceedings of the 6th International Conference on Language Resources and Evaluation, LREC 2008*.

Punjabi to English Bidirectional NMT System

Kamal Deep
Department of Computer Science
Punjabi University, Punjab, India
kamal.1cse@gmail.com

Ajit Kumar
Department of Computer Science
Multani Mal Modi College, Punjab, India
ajit8671@gmail.com

Vishal Goyal
Department of Computer Science
Punjabi University, Punjab, India
vishal.pup@gmail.com

Abstract

Machine Translation is ongoing research for last few decades. Today, Corpus-based Machine Translation systems are very popular. Statistical Machine Translation and Neural Machine Translation are based on the parallel corpus. In this research, the Punjabi to English Bidirectional Neural Machine Translation system is developed. To improve the accuracy of the Neural Machine Translation system, Word Embedding and Byte Pair Encoding is used. The claimed BLEU score is 38.30 for Punjabi to English Neural Machine Translation system and 36.96 for English to Punjabi Neural Machine Translation system.

1 Introduction

Machine Translation (MT) is a popular topic in Natural Language Processing (NLP). MT system takes the source language text as input and translates it into target-language text(Banik et al., 2019). Various approaches have been developed for MT systems, for example, Rule-based, Example-based, Statistical-based, Neural Network-based, and Hybrid-based(Mall and Jaiswal, 2018). Among all these approaches, Statistical-based and Neural Network-based approaches are most popular in the community of MT researchers. Statistical and Neural Network-based approaches are data-driven(Mahata et al., 2018). Both need a parallel corpus for training and validation(Khan Jadoon et al., 2017). Due to this, the accuracy of these systems is higher than the Rule-based system.

The Neural Machine Translation (NMT) is a trending approach these days(Pathak et al., 2018). Deep learning is a fast expanding approach to machine learning and has demonstrated excellent performance when applied to a range of tasks such as speech generation, DNA prediction, NLP, image recognition, and MT, etc. In this NLP tools demonstration, Punjabi to English bidirectional NMT system is showcased.

The NMT system is based on the sequence to sequence architecture. The sequence to sequence architecture converts one sequence into another sequence(Sutskever et al., 2011). For example: in MT sequence to sequence, architecture converts source text (Punjabi) sequence to target text (English) sequence. The NMT system uses the encoder and decoder to convert input text into a fixed-size vector and generates output from this encoded vector. This Encoder-decoder framework is based on the Recurrent Neural Network (RNN)(Wołk and Marasek, 2015)(Goyal and Misra Sharma, 2019). This basic encoder-decoder framework is suitable for short sentences only and does not work well in the case of long sentences. The use of attention mechanisms with the encoder-decoder framework is a solution for that. In the attention mechanism, attention is paid to sub-parts of sentences during translation.

2 Corpus Development

For this demonstration, the Punjabi-English corpus is prepared by collecting from the various online resources. Different processing steps have been done on the corpus to make it clean and useful for the training. The parallel corpus of 259623 sentences is used for training,

Proceedings of the 17th International Conference on Natural Language Processing: System Demonstrations, pages 7–9
Patna, India, December 18 - 21, 2020.

development, and testing the system. This parallel corpus is divided into training (256787 sentences), development (1418 sentences), and testing (1418 sentences) sets after shuffling the whole corpus using python code.

3 Pre-processing of Corpus

Pre-processing is the primary step in the development of the MT system. Various steps have been performed in the pre-processing phase: Tokenization of Punjabi and English text, lowercasing of English text, removing of contraction in English text and cleaning of long sentences (# of tokens more than 40).

4 Methodology

To develop the Punjabi to English Bidirectional NMT system, the OpenNMT toolkit(Klein et al., 2017) is used. OpenNMT is an open-source ecosystem for neural sequence learning and NMT. Two models are developed: one for translation of Punjabi to English and the second for translation of English to Punjabi. The Punjabi vocabulary size of 75332 words and English vocabulary size of 93458 words is developed in the pre-processing step of training the NMT system. For all models, the batch size of 32 and 25 epochs for training is fixed. For the encoder, BiLSTM is used, and LSTM is used for the decoder. The number of hidden layers is set to four in both encode and decoder. The number of units is set to 500 cells for each layer. BPE(Banar et al., 2020) is used to reduce the vocabulary size as the NMT suffers from the fixed vocabulary size. The Punjabi vocabulary size after BPE is 29500 words and English vocabulary size after BPE is 28879 words. “General” is used as an attention function.

By using Python and Flask, a web-based interface is also developed for Punjabi to English bidirectional NMT system. This interface uses the two models at the backend to translate the Punjabi text to English Text and to translate English text to Punjabi text. The user enters input in the given text area and selects the appropriate NMT model from the dropdown and then clicks on the submit button. The input is pre-processed, and then the NMT model translates the text into the target text.

Model	BLEU score
Punjabi to English NMT model	38.30
English to Punjabi NMT model	36.96

Table 1: BLEU score of both models

5 Results

Both proposed models are evaluated by using the BLEU score(Snover et al., 2006). The BLEU score obtained at all epochs is recorded in a table for both models. Table 1 shows the BLEU score of both models. The best BLEU sore claimed is 38.30 for Punjabi to English Neural Machine Translation system and 36.96 for English to Punjabi Neural Machine Translation system.

References

Nikolay Banar, Walter Daelemans, and Mike Kestemont. 2020. Character-level Transformer-based Neural Machine Translation, arXiv: 2005.11239.

Debajyoty Banik, Asif Ekbal, Pushpak Bhattacharyya, Siddhartha Bhattacharyya, and Jan Platos. 2019. Statistical-based system combination approach to gain advantages over different machine translation systems. *Heliyon*, 5(9):e02504.

Vikrant Goyal and Dipti Misra Sharma. 2019. LTRC-MT Simple & Effective Hindi-English Neural Machine Translation Systems at WAT 2019. In *Proceedings of the 6th Workshop on Asian Translation,Hong Kong, China*, pages 137–140.

Nadeem Khan Jadoon, Waqas Anwar, Usama Ijaz Bajwa, and Farooq Ahmad. 2017. Statistical machine translation of Indian languages: a survey. *Neural Computing and Applications*, 31(7):2455–2467.

Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, Alexander M. Rush, Josep Crego, Jean Senellart, and Alexander M. Rush. 2017. OpenNMT: Open-source Toolkit for Neural Machine Translation. *ACL 2017 - 55th Annual Meeting of the Association for Computational Linguistics, Proceedings of System Demonstrations*:67–72.

Sainik Kumar Mahata, Soumil Mandal, Dipankar Das, and Sivaji Bandyopadhyay. 2018. SMT vs NMT: A Comparison over Hindi & Bengali Simple Sentences. In *International Conference on Natural Language Processing*, number December, pages 175–182.

Shachi Mall and Umesh Chandra Jaiswal. 2018. Survey: Machine Translation for Indian Language. *International Journal of Applied Engineering Research*, 13(1):202–209.

Amarnath Pathak, Partha Pakray, and Jereemi Bentham. 2018. English–Mizo Machine Translation using neural and statistical approaches. *Neural Computing and Applications*, 31(11):7615–7631.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. *AMTA 2006 - Proceedings of the 7th Conference of the Association for Machine Translation of the Americas: Visions for the Future of Machine Translation*:223–231.

Ilya Sutskever, James Martens, and Geoffrey Hinton. 2011. Generating Text with Recurrent Neural Networks. *Proceedings of the 28th International Conference on Machine Learning*, 131(1):1017–1024.

Krzysztof Wołk and Krzysztof Marasek. 2015. Neural-based Machine Translation for Medical Text Domain. Based on European Medicines Agency Leaflet Texts. *International Conference on Project MANagement*, 64:2–9.

Software to Extract Parallel Data from English-Punjabi Comparable Corpora

Manpreet Singh Lehal[1*], Dr. Ajit Kumar[2], Dr. Vishal Goyal[3]
[1]Department of Computer Science, Lyallpur Khalsa College, Jalandhar
[2]Associate Professor, Department of Computer Science, Multani Mal Modi College, Patiala
[3]Department of Computer Science, Punjabi University, Patiala
Email: mslehal@lkc.ac.in {ajit8671, vishal.pup}@gmail.com

Abstract

Machine translation from English to Indian languages is always a difficult task due to the unavailability of a good quality corpus and morphological richness in the Indian languages. For a system to produce better translations, the size of the corpus should be huge. We have employed three similarity and distance measures for the research and developed a software to extract parallel data from comparable corpora automatically with high precision using minimal resources. The software works upon four algorithms. The three algorithms have been used for finding Cosine Similarity, Euclidean Distance Similarity and Jaccard Similarity. The fourth algorithm is to integrate the outputs of the three algorithms in order to improve the efficiency of the system.

1. Introduction

Machine translation from English to Indian languages is always a difficult task due to the unavailability of a good quality corpus and morphological richness in the Indian languages. For a system to produce better translations, the size of the corpus should be huge. In addition to that, the parallel sentences should convey similar meanings, and the sentences should cover different domains. Modelling the system with such a corpus can assure good translations while testing the model. Since English - Punjabi language pair is an under-resourced pair, this study provides a breakthrough in acquiring English - Punjabi Corpus for performing the task of machine translation. We have employed Statistical methods for the research and developed a software to extract parallel data from comparable corpora automatically with high precision using minimal resources.

We generate an English-Punjabi Comparable Corpora which is used as input data. We have used the articles from Wikipedia which are stored in the dump. The articles of English and Punjabi languages are extracted, aligned and refined. We also received access to the database of Indian Language Technology Proliferation and Deployment Centre (TDIL) and used the noisy parallel sentences. Sentences were also collected from Gyan Nidhi corpus and reports of college activities. Thus, our data is not restricted to one particular domain.

We employ three similarity measures of Cosine Similarity, Jaccard Distance and Euclidean Distance to find the similarity of two English Corpora. Firstly, the algorithms are performed individually and then the integrated approach is used by combining the results of all the three similarity measuring algorithms to reach better output levels. The software works upon four algorithms. The three algorithms have been used for finding Cosine Similarity, Euclidean Distance Similarity and Jaccard Similarity. The fourth algorithm is to integrate the outputs of the three algorithms in order to improve the efficiency of the system. The codes for similarity algorithms have been implemented in python using Scikit Learn. The sentences are first converted into vectors using tf-idf vectorization and then the algorithms are employed.

Every similarity measure has its own limitations when used individually. Combining the scores of three similarity measures complements the features and give better results. Only those

Proceedings of the 17th International Conference on Natural Language Processing: System Demonstrations, pages 10–12
Patna, India, December 18 - 21, 2020.

translation pairs are selected which are similarly paired in all the three algorithms. The translation pairs which do not occur in the output of one or two algorithms are discarded.

We run the three similarity algorithms and obtain similarity scores. Threshold values are fixed by getting the average of all the similarity scores obtained for each algorithm. In case of Euclidean Distance and Jaccard Distance, the translation pairs having similarity scores below the threshold values are selected and in case of Cosine similarity translation pairs with similarity scores above the threshold value are selected. The remaining translation pairs are filtered out. This refines the output to a great extent.

The results obtained from the three algorithms are integrated aiming at the improvement of the output translation pairs and finding the best pairs. Only those translation pairs are selected which are similarly paired in all the three algorithms. The translation pairs which do not occur in the output of one or two algorithms are discarded.

With the integrated approach, we are able to achieve a precision level of 93 percent and accuracy is 86 percent. The results make it clear that the integrated approach improve the results to a great extent and thus, validate the usage of this approach.

There are three components of the web interface of the software: Punjabi Input, English Input and Aligned Data. The input data in form of sentences or paragraphs is copied in relevant language boxes on the left side and submitted. It gives the translation pair output in the Aligned data box on the right-side. The tuning button is used to identify translations at the required level of similarity. It can be increased or decreased to find exact parallel sentences as well as translations pairs similar at phrase level.

References

Afli, H., Barrault, L. and Schwenk, H. (2014) 'Multimodal Comparable Corpora for Machine Translation'.

Artetxe, M. and Schwenk, H. (2019) 'Massively Multilingual Sentence Embeddings for Zero-Shot Cross-Lingual Transfer and Beyond', Transactions of the Association for Computational Linguistics. doi: 10.1162/tacl_a_00288.

Bharadwaj, R. G. and Varma, V. (2011) 'Language independent identification of parallel sentences using Wikipedia', Proceedings of the 20th international conference companion on World wide web - WWW '11, p. 11. doi: 10.1145/1963192.1963199.

Cettolo, M., Federico, M. and Bertoldi, N. (2010) 'Mining Parallel Fragments from Comparable Texts', Iwslt-2010, pp. 227–234.

Chu, C., Nakazawa, T. and Kurohashi, S. (2013) 'Chinese – Japanese Parallel Sentence Extraction from Quasi – Comparable Corpora', pp. 34–42.

Deep, K., Kumar, A. and Goyal, V. (2018) 'Development of Punjabi-English (PunEng) Parallel Corpus for Machine Translation System', International Journal of Engineering & Technology. doi: 10.14419/ijet.v7i2.10762.

Dwivedi S.K , Sukhadeve, P. P. (2010) 'Machine Translation System in Indian Perspectives', 6(10), pp. 1082–1087.

Dzmitry, B., Kyunghyun, C. and Yoshua, B. (2014) 'Neural machine translation by jointly learning to align and translate', 3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings.

Eisele, A. and Xu, J. (2010) 'Improving Machine Translation Performance Using Comparable Corpora', in Proceedings of the 3rd Workshop on Building and Using Comparable Corpora. Workshop on Building and Using Comparable Corpora (BUCC-3), Applications of Parallel and Comparable Corpora in Natural Language Engineering and the Humanities, La Valletta, Malta, pp. 35–41.

Fu, X. et al. (2013) 'Phrase-based Parallel Fragments Extraction from Comparable Corpora', Proceedings of the Sixth International Joint Conference on Natural Language Processing, (October), pp. 972–976. Available at: http://aclweb.org/anthology/I13-1129.

Fung, P. et al. (2004) 'Mining Very-Non-Parallel Corpora: Parallel Sentence and Lexicon Extraction via Bootstrapping and EM', EMNLP 2004 - Conference on Empirical Methods in Natural Language Processing, pp. 57–63. Available at: http://www.aclweb.org/anthology-new/W/W04/W04-3208.pdf.

Goyal, V., Kumar, A. and Lehal, M. S. (2020) 'Document Alignment for Generation of English-Punjabi Comparable Corpora from Wikipedia', International Journal of E-Adoption. doi: 10.4018/ijea.2020010104.

HEWAVITHARANA, S. and VOGEL, S. (2016) 'Extracting parallel phrases from comparable data

for machine translation', Natural Language Engineering. doi: 10.1017/s1351324916000139.

Jindal, S., Goyal, V. and Singh, J. (2017) 'Building English-Punjabi Parallel corpus for Machine Translation', International Journal of Computer Applications. doi: 10.5120/ijca2017916036.

Kumar, A. and Goyal, V. (2018) 'Hindi to Punjabi machine translation system based on statistical approach', Journal of Statistics and Management Systems. Taylor & Francis, 21(4), pp. 547–552.

Kvapil\'\iková, I. et al. (2020) 'Unsupervised Multilingual Sentence Embeddings for Parallel Corpus Mining', in Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop, pp. 255–262.

Lehal, M. S., Kumar, A. and Goyal, V. (2018) 'Review of techniques for extraction of bilingual lexicon from comparable corpora', International Journal of Engineering and Technology(UAE). doi: 10.14419/ijet.v7i2.30.13456.

Lehal, M. S., Kumar, A. and Goyal, V. (2019) 'Comparative analysis of similarity measures for extraction of parallel data', International Journal of Control and Automation.

Munteanu, D. M. D. (2002) 'Processing comparable corpora with Bilingual Suffix Trees', Proceedings of the ACL-02 conference on Empirical methods in natural language processing, 10, pp. 289–295.

Pascale Fung, E. P. and S. S. (2010) 'Trillions of comparable documents', Proceedings of the 3rd workshop on building and using comparable corpora: from parallel to non-parallel corpora, (May), pp. 26–34.

Quirk, C., Udupa, R. and Menezes, A. (2007) 'Generative Models of Noisy Translations with Applications to Parallel Fragment Extraction', Machine Translation Summit XI, (2000). Available at: http://www.mt-archive.info/MTS-2007-Quirk.pdf.

Rauf, S. A. and Schwenk, H. (2011) 'Parallel sentence generation from comparable corpora for improved SMT', Machine translation. Springer, 25(4), pp. 341–375.

Riesa, J. and Marcu, D. (2012) 'Automatic Parallel Fragment Extraction from Noisy Data', in Proc. NAACL, pp. 538–542.

Tillmann, C. (2009) 'A Beam-Search Extraction Algorithm for Comparable Data', Acl-2009, (August), p. 4. doi: 10.3115/1667583.1667653

A Sanskrit to Hindi Language Machine Translator using Rule Based Method

Prateek Agrawal
Lovely Professional University,
Punjab, India.
Univeristy of Klagenfurt, Austria
prateek061186@gmail.com

Vishu Madaan
Lovely Professional University,
Punjab, India.
vishumadaan123@gmail.com

1 Demonstration

Hindi and Sanskrit both the languages are having the same script i.e. Devnagari Script which results in few basic similarities in their grammar rules. As we know that Hindi ranks fourth in terms of speaker's size in the world and over 60 Million people in India are Hindi internet users [1]. In India itself, there are approximately 120 languages and 240 mother tongues but hardly a few languages are recognized worldwide while the others are losing their existence in society day by day. Likewise, Sanskrit is one among those important languages that are being ignored in society. As per census report of India in 2001, less than 15000 citizens have returned Sanskrit as their Mother tongue or preferred medium of communication. A key reason behind poor acceptance of Sanskrit is due to language barrier among Indian masses and lack of knowledge about this language among people. Therefore, our attempt is just to connect a big crowd of Hindi users with Sanskrit language and make them familiar at least with the basics of Sanskrit. We developed a translation tool that parses Sanskrit words (prose) one by one and translate it into equivalent Hindi language in step by step manner:

- We created a strong Hindi-Sanskrit corpus that can deal with Sanskrit words effectively and efficiently (Agrawal and Jain, 2019).
- We proposed an algorithm to stem Sanskrit word that chops off the starts / ends of words to find the root words in the form of nouns and verbs (Jain and Agrawal, 2015).
- After stemming, we developed an algorithm to search the equivalent Hindi meaning of stemmed words from the corpus based on semantic analysis (Bhadwal et al., 2019).
- We developed an algorithm to implement semantic analysis to translate words that helps the tool to identify required parameter details like gender, number, case etc.
- Next, we developed an algorithm for discourse integration to disjoin each translated sentence based on subject / noun dependency (Bhadwal et al., 2020).
- Next, we implemented pragmatic analysis algorithm that ensures the meaningful validation of these translated Hindi sentences syntactically and semantically (Agrawal and Jain, 2019).
- We further extended our work to summarize the translated text story and suggest the suitable heading / title. For this, we referred ripple down rule-based parts of speech (RDR-POS) Tagger (Dat et al., 2016) for word tagging in the POS tagger corpora.
- We proposed a title generation algorithm which suggests some suitable title of translated text (Jain and Agrawal, 2018).
- Finally, we assembled all phases to one translation tool that takes a story of maximum one hundred words as input and translates it into equivalent Hindi language.

Figure 1 shows the working demonstration of proposed translation tool in nine window panels and 18 widgets (objects) to describe each step. In the first window panel, 3 objects are highlighted and numbered as 1, 4 and 5. Users can type any Sanskrit sentence using a simple QWERTY keyboard in the text area (highlighted as Object 1).

[1] http://www.business-standard.com/article/current-affairs/hindi-internet-users-estimated-at-60-million-in-india-survey-116020400922$_1$.*html*

Proceedings of the 17th International Conference on Natural Language Processing: System Demonstrations, pages 13–15
Patna, India, December 18 - 21, 2020.

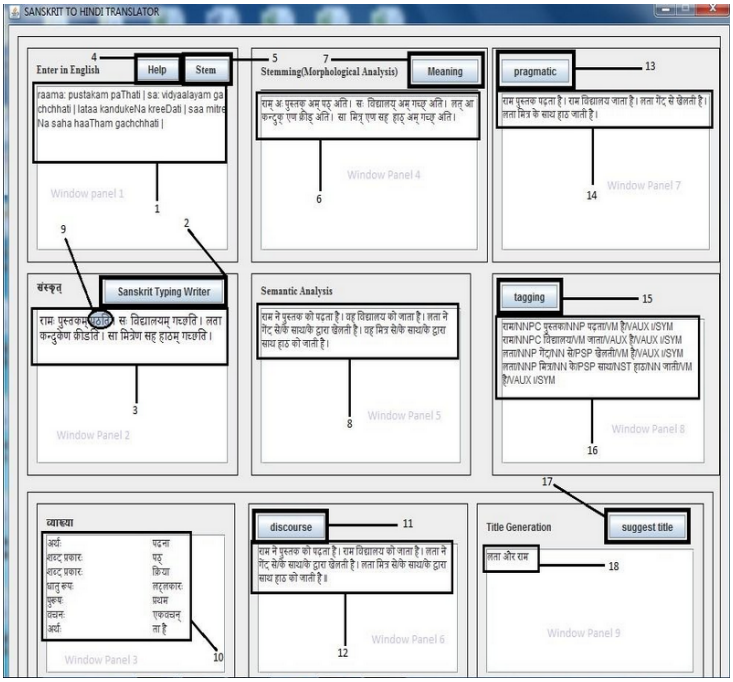

Figure 1: Screenshot of Output Interface to translate Sanskrit sentence into Hindi Language

Help button (as Object 4), is added to help novice users to type Sanskrit sentence(s) in English. Simultaneously, transliterated sentence from English to Sanskrit is displayed in the second window panel as object 3. Sentences can also be typed using a virtual keyboard (highlighted as object 2) [6]. Next step is to stem Sanskrit words written in the 2nd window panel by pressing. Stem button (highlighted as object 5 in first window panel). Output of the same is displayed in the 4th window panel and highlighted as object 6. Next step is to press the button "Meaning" (Object 7) to get Hindi meaning of stemmed Sanskrit words (highlighted as Object 8 in window panel 5).

Once we get raw or intermediate meaning of Sanskrit words into Hindi as object 8, we click on the Discourse button (as object 11) to get the translated results in discourse form (highlighted as object 12 in window panel 6). As an option for easy understanding, our system has been felicitated to explain the meaning of Sanskrit words into Hindi with complete explanation (highlighted as object 9 and 10). Once Discourse integration is done, the next step is to get the correct meaning of sentences (as object 14) by clicking on the pragmatic button (as object 13) in window panel 7. Final translated sentences are passed through RDR POS tagger for tagging purposes (as highlighted in window panel 8 using objects 15 and 16). Finally, the system implements an algorithm to generate appropriate title(s) of translated sentences in Hindi using object button 17 in Window panel 9 and displays the result as object 18.

References

Prateek Agrawal and Leena Jain. 2019. Anuvaadika: Implementation of sanskrit to hindi translation tool using rule-based approach. *Recent Patents on Computer Science*, 12(2):10–21.

Neha Bhadwal, Prateek Agrawal, and Vishu Madaan. 2019. Bilingual machine translation system between hindi and sanskrit languages. In *International Conference on Advanced Informatics for Computing Research*, pages 312–321. https://link.springer.com/chapter/10.1007/978-981-15-0108-1_29.

Neha Bhadwal, Prateek Agrawal, and Vishu Madaan. 2020. A machine translation system from hindi to sanskrit language using rule based approach. *Scalable Computing: Practice and Experience*, 21(3):543–553.

Quoc Nguyen Dat, Quoc Nguyen Dai, Duc Pham Dang, and Bao Pham Son. 2016. A robust transformation-based learning approach using ripple down rules for part-of-speech tagging. *AI Communications (AICom)*, 29(3):409–422.

Leena Jain and Prateek Agrawal. 2015. Text independent root word identification in hindi language using natural language processing. *International Journal of Advanced Intelligence Paradigms*, 7(3/4):240.

Leena Jain and Prateek Agrawal. 2018. Sheershak: an automatic title generation tool for hindi short stories. In *2018 International Conference on Advances in Computing, Communication Control and Networking (ICACCCN)*, pages 579–584. IEEE.

Urdu To Punjabi Machine Translation System

Umrinderpal Singh[1] , Vishal Goyal[2] , Gurpreet Singh Lehal[3]
Department of Computer Science GHG Khalsa College Gurusar Sadhar Ludhiana India[1],
Department of Computer Science, Punjabi University Patiala India[2,3]
{umrinderpal[1], vishal.pup[2], gslehal[3]}@gmail.com

Abstract

Machine Translation is a popular area of NLP research field. There are various approaches to develop a machine translation system like Rule-Based, Statistical, Neural and Hybrid. A rule-Based system is based on grammatical rules and uses bilingual lexicons. Statistical and Neural use the large parallel corpus for training the respective models. Where the Hybrid MT system is a mixture of different approaches. In these days the corpus-based machine translation system is quite popular in NLP research area. But these models demands huge parallel corpus. In this research, we have used a hybrid approach to develop Urdu to Punjabi machine translation system. In the developed system, statistical and various sub-system based on the linguistic rule has been used. The system yield 80% accuracy on a different set of the sentence related to domains like Political, Entertainment, Tourism, Sports and Health. The complete system has been developed in a C#.NET programming language.

1. About the system

In this digital era, where different communities across the world are interacting with each other and sharing digital resources. In such kind of digital world natural languages are obstacles in communication. To remove this obstacle from communication, NLP researchers are working to develop Machine Translation systems. These Machine Translation systems can detect various languages and their domains and automatically translate source language text to target-language text. The machine translation system can be developed using various approaches, for example, Rule-Based, Example-Based, Statistical, Neural and various hybrid approaches(Antony P.J 2013). The Rule-based and example-based systems are based on various linguistic rules and a large lexicons dictionaries (Goyal V and G S Lehal 2010). These dictionaries contained parallel word and phrases of source and target language. The statistical system is purely based on some statistical model. These systems required huge parallel corpus to train the model(G S Josan and G S Lehal 2008). The system automatically creates parallel dictionaries and phrase tables from a given parallel corpus (Goyal V and G S Lehal 2010). In this approach researcher's main task is to create or arrange a parallel corpus. Most of the other work simply was given to the machine-like creating phase table and learning the model etc based on the parallel corpus. The neural machine translation is a trending approach these days. In a neural-based approach, deep learning is fast expanding approach for Machine Translation and many other research areas of computing. This approach required a large parallel corpus to train the system and other tasks like creating and learning the translation rules are automatically handled by training algorithms. The Statistical and Neural Machine Translation system can yield excellent results but required a huge parallel corpus for training (Ajit Kumar and Vishal Goyal

Proceedings of the 17th International Conference on Natural Language Processing: System Demonstrations, pages 16–18
Patna, India, December 18 - 21, 2020.

2011). There are many languages in the world which are resource-poor, they don't have any large enough corpus to train the statistical and neural-based system. Urdu and Punjabi languages are one of them. To the best of our knowledge, there is no large enough parallel corpus is available for Urdu and Punjabi language pair to train the statistical and neural-based model. Along with this, Urdu and Punjabi are morphological rich languages. These languages required many other resources like stemmer (Lehal, G. 2009)., lemmatizers, transliteration, spell checker, grammar checker, font detection and conversion tools.

For this demonstration, the system has been developed to translate Urdu to Punjabi Unicode text. As mentioned previously, Urdu and Punjabi are resource-poor languages therefore we have developed various preprocessing, translation and post-processing tools to refine the translated text. In the preprocessing phase, we have developed sentence and word tokenization models for Urdu. The word tokenization system for Urdu is not simple like any other language. In Urdu, segmentation issue (Lehal, Gurpreet Singh. 2010) is the key challenge therefore the sub-system has been developed to handle this issue in preprocessing. In the preprocessing phase, the text classification system has been developed to classify the input Urdu text into various predefined domains like Political, Entertainment, Health and Tourism. The Naive Bayes approached has been used for the text classification system. The reason to use the text classification system to apply the specific knowledge-based on the given input text. By using this clarification module the system can remove various ambiguities in translation. The system used the Hidden Markov Model as a learning module and the Viterbi algorithm has been used as a decoder. Urdu and Punjabi are closely related languages and share grammatical structure and word order. Therefore the system does not require a word reordering module. The system takes manually mapped words and phrases to generate translation probabilities. The language model for translation has been developed using Kneser-Ney smoothing algorithm. Along with the preprocessing, training and translation modules, various sub-system has been developed to refine the output, for example, removing diacritical marks, Izafaat word checking, stemming and creating inflations and transliteration sub-system to handle unknown words. The text classification system's overall accuracy is 96% and complete translation system's accuracy is more than 80% on various domains. The system mainly trained and tested for Political, Sports, Entertainment, Tourism and Health domains. The training data has been collected from BCC Urdu website and TDIL 50000 thousand parallel sentences. On average the system knowledge-based for different domains contains 56023 mapped phrases and words. The system's phrase table incorporates a maximum length of the phrase was four-gram. The total 2088 four-gram phrase used in phrase table. In the translation process, most of the time uni-gram, bi-gram and tri-gram phrases were sufficient to translate any given Urdu input text. The working system is available on http://u2p.learnpunjabi.org/ URL.

References

Ajit Kumar and Vishal Goyal (2011) "Comparative Analysis of Tools Available for Developing Statistical Approach Based Machine Translation System", ICSIL 2011 CCIS 139, pp: 254-260

Antony P.J (2013) "Machine Translation Approachesand Survey for Indian Languages" Computationa Linguistics and Chinese Language Processing Vol.18, No. 1, March 2013, pp: 47-78

G S Josan and G S Lehal (2008), "A Punjabi to Hindi Machine Translation System" in proceeding Coling: Companion volume: Posters and Demonstrations, Manchester, UK, pp: 157-160

Goyal V and G S Lehal (2010), "Web based Hindi to Punjabi machine translation system", J. Emerg. Technol. Web Intell, pp: 148-151

Goyal, V., & Lehal, G. S. (2009). "Evaluation of Hindi to Punjabi Machine Translation System". IJCSI

International Journal of Computer Science, 4(1), 36- 39.

G.S Josan and G.S Lehal "Evaluation of Direct Machine Translation System For Punjabi To Hindi" http://www.learnpunjabi.org/pdf/directtrans.pdf Accessed on: 1/12/2020

Josan G S and G S Lehal (2008) “Punjabi to Hindi machine translation system”, in Proceedings of the 22nd International Conference on Computational Linguistics, MT-Archive, Manchester, UK, pp: 157-160

Lehal, G S. (2010) “A word segmentation system for handling space omission problem in Urdu script”. 23rd International Conference on Computational Linguistics.

Lehal, G. (2009). “A Two Stage Word Segmentation System for Handling Space Insertion Problem in Urdu Script”, World Academy of Science, Engineering and Technology 60.

The Hindi to Dogri Machine Translation System

Preeti Dubey
Department of Computer Applications,
Govt. College for Women, Parade Ground, Jammu
{preetidubey2000}@yahoo.com

Abstract

The Hindi to Dogri Machine translation system is a rule-based MT developed and copyrighted by GoI in 2014. It is the first system developed to convert Hindi text into Dogri (the regional language of Jammu). The system is developed using ASP.Net and the databases are in MS-Access. This Machine Translation system accepts Hindi text as input and provides Dogri text as output in Unicode.

Working of the System: The brief working of the system is discussed in this section. The system has a very easy to use and simple interface. The Hindi text to be converted into Dogri is placed in a textbox and the user needs to click on the translate button to get the output in another text box as shown below in Fig1:

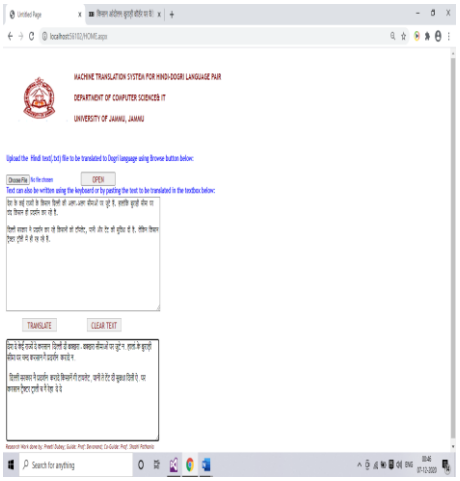

Fig1: shows a screenshot of the Hindi to Dogri Machine Translation System

The overall accuracy of Hindi to Dogri Machine Translation system is found to be approximately 98 %. In case of technical text and news, the output accuracy is tested to be 98% and above whereas incase of stories the accuracy level falls due to use of ambiguous words in stories. The system handles ambiguity of the hindi words "और" and "से". The system performance can be improved by incorporating word sense disambiguation of more ambiguous words.

Some Text translated using this system is shown below in Table 1 :

S.No	Hindi Input	Dogri Output
1.	दिल्ली सरकार ने प्रदर्शन कर रहे किसानों को टॉयलेट, पानी और टेंट की सुविधा दी है. लेकिन किसान ट्रैक्टर ट्रॉली में ही रह रहे हैं.	दिल्ली सरकार नै प्रदर्शन करादे किसानें गी टायलेट , पानी ते टेंट दी सुबधा दित्ती ऐ . पर करसान ट्रैक्टर ट्राली च गै रेहा दे दे
2.	डॉ.भीमराव आंबेडकर ने एक इंटरव्यू में कहा था कि महात्मा गांधी सिर्फ इतिहास का हिस्सा हैं, युग निर्माण करने वाले नहीं.	डा भीमराव आंबेडियै नै इक इंटरव्यू च आखेआ हा जे महात्मा गांधी सिर्फ इतिहास दा हिस्सा न , जुग निरमान करने आह्ले नेईं .
3.	संसद भवन की नई बिल्डिंग का भूमि पूजन 10 दिसंबर को होने वाला है.	संसद भवन दी नमीं बिल्डिंग दा जमीन पूजन 10 दिसंबर गी होने आह्ला ऐ .

Table1: shows text translated using the Hindi to Dogri Machine Translation System

Proceedings of the 17th International Conference on Natural Language Processing: System Demonstrations, pages 19–20
Patna, India, December 18 - 21, 2020.

The incorrect translations in each text are kept bold. The accuracy of the system is clearly visible in the above translated text to be above 98%. More text translations will be discussed during the demonstration of the system.

Opinion Mining System for Processing Hindi Text for Home Remedies Domain

Arpana Prasad
Department of Computer Science, Punjabi University, Patiala.
arpanaprasad2013@gmail.com

Neeraj Sharma
Department of Computer Science, Punjabi University, Patiala.
sharma_neeraj@hotmail.com

Shubhangi Sharma
SAP Labs India Ltd.
shubhangi.sharma@sap.com

Abstract

Lexical and computational components developed for an Opinion Mining System that process Hindi text taken from weblogs are presented in the paper. Text chosen for processing are the ones demonstrating cause and effect relationship between related entities 'Food' and 'Health Issues'. The work is novel and lexical resources developed are useful in the current research and may be of importance for future research.

1 Introduction

Opinion Mining (OM) is a field of study in Computer Science that deals with development of software applications related to text classifications and summarizations. Researchers working in this field contribute lexical resources, computing methodologies, text classification approaches, and summarization modules to perform OM tasks across various domains and different languages. The common challenges encountered by the researchers in this field are; domain dependency, classification of sarcastic text, dealing with comments with thwarted expectations, language dependency, extensive domain knowledge, and categorising of text as a fact bearing text or an opinion bearing fact. There has been development in this field for processing Hindi unstructured text(Arora, 2013; Arora, Bakliwal, & Varma, 2012; Harikrishna & Rao, 2019; Jha, Savitha, Shenoy, Venugopal, & Sangaiah, 2017; Joshi, R, & Bhattacharyya, 2010; Mishra, Venugopalan, & Gupta, 2016; Mogadala & Varma, 2012; Reddy & Sharoff, 2011; Sharma, Nigam, & Jain, 2014) .

2 Research Problem

An ongoing research in the field of OM, which is dedicated towards the development of lexical and software components that facilitate opinion classification and summarization from Hindi Text appearing on Web logs is presented here. Hindi text showcasing cause and effect relationship between 'Food' and 'Health Issues' entities are processed in the OMS under study. The resources are developed for an algorithm 'A' such that for a sentence 'Y' which is a domain specific sentence from weblogs in Hindi, A(Y) returns a set {F, HI, p, s} such that F is a subset of set, FOOD={set of word or phrases in Hindi used for an edible item and HI is a subset of set, HEALTH_ISSUE= {{set of word or phrases in Hindi used for a part of body composition 'BODY_COMPONENT'} UNION {set of word or phrases in Hindi used for a health problem a human being face 'HEALTH_PROBLEM'}}. Element 'p' takes numeric value '1' or '-1' where value '1' means that from the text 'Y', algorithm 'A' computationally derived that the food entities mentioned in set 'F' have a positive effect in health issues mentioned in set 'HI' and the numeric value '-1' means that the food entities in set 'F' have a negative effect in health issues in set 'HI'. The element's' may take value '1' or '2' indicating that the strength of polarity 'p' is medium or strong. This research is undertaken with a sole objective to contribute towards bringing text appearing in Hindi weblogs under the benefits that an OM system offers. Language dependent computational contributions based on English and Chinese languages, in the previous studies motivated the proposal of the current work (Bhattacharya, 2014; Miao, Zhang, Zhang, & Yu, 2012; Yang, Swaminathan, Sharma, Ketkar, & D'Silva, 2011). This research is undertaken with a sole objective to contribute towards bringing text appearing in Hindi

Proceedings of the 17th International Conference on Natural Language Processing: System Demonstrations, pages 21–23
Patna, India, December 18 - 21, 2020.

weblogs under the benefits that an OM system offers.

3 Key Issues Identified and Addressed

Lexical resources required for Named Entity Recognition (NER) and polarity classification were are not available for the current research. The same are developed in the research. A corpus in same domain and in same language as the text being processed in an OM system helps in devising computational components for the system and evaluating them. An annotated corpus of Hindi text relevant to the research did not pre-exist hence it is developed in the research.

4 Major Contributions

A domain specific Hindi corpus, with semantic and syntactic annotations is developed in the research. The semantic annotations of the corpus help in devising and evaluating the algorithms formulated in the research. A total of 3303 unique domain specific Hindi sentences from weblogs are collected for the corpus. The corpus has approximately 2516 unique sentences with positive annotations for 'FOOD' and 'HEALTH ISSUE' associations and 787 unique sentences with negative annotations for associations between entities. The total number of syntactically annotated Hindi words in the corpus is 60000 approximately. Domain specific lexical datasets for entities; FOOD, FOOD_ADJECTIVE, FOOD_COMPONENT, BODY_COMPONENT and HEALTH_PROBLEM are also developed/identified in the research. The total number of Hindi words/phrases identified for the lexical datasets are 8000 lexicons approximately. A lexical based polarity classification and polarity strength classification algorithm is developed in the research. There are two datasets developed to support the algorithm. The datasets are; (a) a set of approx. 15000 positive polarity bearing phrases and a set of approx. 10000 negative polarity bearing phrases that generate a vocabulary of approximately 35000 trigram words, (b) a set of approximately 14000 strength determining phrases that are useful in determining the strength of the polarity identified by the algorithm.

5 Methodology Adopted

Firstly, a set of domain specific keywords were consolidated. Then for approximately 2 years using some identified keywords domain specific Hindi text from weblogs were collected. Semi automated processes were adopted for data cleaning and refinement of the collected sentences for the corpus. The corpus is syntactically annotated with part of speech (POS) using a POS tool for Hindi developed by CFILT, IIT Bombay. Each word of the sentence is stemmed to give flexibility to the word usage for NER. The corpus is semantically annotated with the help of two annotators. A detailed guideline for semantic annotations was developed in the research. The annotators were supposed to annotate the text on their perception about the related entities and polarity of associations of the related entities. Domain specific lexical datasets are developed using Hindi WordNet of CFILT, IIT Bombay. The developed lexical datasets are used by the NER algorithm that extracts domain specific related entities from text under OM processing. The polarity bearing and strength determining phrases are developed using a seed list from the corpus, later the dataset was extended using phrases from books that are candidate for appearing on weblogs and using synonyms of words in phrases with same POS using Hindi WordNet. Finally, a lexical based algorithm using Naive Bayes Classifier that is trained on a vocabulary of n-gram words from polarity phrases and strength determining phrases is formalized for polarity classification of association and strength classification of polarity.

6. Experimental Results

All the lexical resources are developed using SQLite Studio 3.1.1. The algorithms are developed using Python 3.7. The classification algorithms when trained using trigram words from polarity bearing phrases and strength determining phrases and tested on a random set of 900 sentences from the corpus gives best results with Accuracy: 0.996, Precision: 0.998 Recall: 0.994 and F-Score of 0.996. The named entity recognition algorithm tested on the same dataset gives 85% accuracy. The lexical outcomes of this research may be useful to other researchers working in related fields.

References

Arora, P. (2013). Sentiment Analysis for Hindi Language. International Institute of Information Technology. Retrieved from coling2017.pdf (iiit.ac.in)

Arora, P., Bakliwal, A., & Varma, V. (2012). Hindi Subjective Lexicon Generation using WordNet Graph Traversal. International Journal of Computational Linguistics and Applications, 3(Jan-Jun 2012), 25–29.

Bhattacharya, S. (2014). Computational methods for mining health communications in web 2. 0. University of Iowa. Retrieved from http://ir.uiowa.edu/etd/4576

Harikrishna, D. M., & Rao, K. S. (2019). Children's Story Classification in Indian Languages Using Linguistic and Keyword-based Features. ACM Transactions on Asian and Low-Resource Language Information Processing, 19(2). https://doi.org/https://doi.org/10.1145/3342356

Jha, V., Savitha, R., Shenoy, P. D., Venugopal, K. R., & Sangaiah, A. K. (2017). A novel sentiment aware dictionary for multi-domain sentiment classification, 0, 1–13. https://doi.org/10.1016/j.compeleceng.2017.10.015

Joshi, A., R, B. A., & Bhattacharyya, P. (2010). A Fall-back Strategy for Sentiment Analysis in Hindi: a Case Study. In Proceedings of the ICON-2010:8th International Conference on Natural Language Processing. Macmillan Publishers, India. Retrieved from http://ltrc.iiit.ac.in/proceedings/ICON-2010

Miao, Q., Zhang, S., Zhang, B., & Yu, H. (2012). Extracting and Visualizing Semantic Relationships from Chinese Biomedical Text. Proceedings of the 26th Pacific Asia Conference on Language, Information, and Computation, 99–107. Retrieved from http://aclweb.org/anthology/Y12-1010

Mishra, D., Venugopalan, M., & Gupta, D. (2016). Context Specific Lexicon for Hindi Reviews. Procedia Computer Science, 93(September), 554–563. https://doi.org/10.1016/j.procs.2016.07.283

Mogadala, A., & Varma, V. (2012). Retrieval approach to extract opinions about people from resource scarce language news articles. Proceedings of the First International Workshop on Issues of Sentiment Discovery and Opinion Mining - WISDOM '12, (August), 1–8. https://doi.org/10.1145/2346676.2346680

Reddy, S., & Sharoff, S. (2011). Cross language POS taggers (and other tools) for Indian languages: An experiment with Kannada using Telugu resources. Cross Lingual Information Access, 11–19. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.222.1557&rep=rep1&type=pdf#page=27

Sharma, R., Nigam, S., & Jain, R. (2014). Polarity Detection of Movie Reviews in Hindi Language. International Journal on Computational Science & Applications, 4(4), 49–57. https://doi.org/10.5121/ijcsa.2014.4405

Yang, H., Swaminathan, R., Sharma, A., Ketkar, V., & D'Silva, J. (2011). Mining biomedical text towards building a quantitative food-disease-gene network. Studies in Computational Intelligence, 375, 205–225. https://doi.org/10.1007/978-3-642-22913-8_10arXiv:1503.06733. Version 2

Sentiment Analysis of English-Punjabi Code-Mixed Social Media Content

Mukhtiar Singh[1], Vishal Goyal[2], Sahil Raj[3]
[1,2]Department of Computer Science, Punjabi University, Patiala
[3]School of Management and Studies, Punjabi University, Patiala
{mukhtiarrai73,vishal.pup,dr.sahilraj47}@gmail.com

Abstract

Sentiment analysis is a field of study for analyzing people's emotions, such as Nice, Happy, ਦੁਖੀ (sad), changa (Good), etc. towards the entities and attributes expressed in written text. It noticed that, on microblogging websites (Facebook, YouTube, Twitter), most people used more than one language to express their emotions. The change of one language to another language within the same written text is called code-mixing. In this research, we gathered the English-Punjabi code-mixed corpus from micro-blogging websites. We have performed language identification of code-mix text, which includes Phonetic Typing, Abbreviation, Wordplay, Intentionally misspelled words and Slang words. Then we performed tokenization of English and Punjabi language words consisting of different spellings. Then we performed sentiment analysis based on the above text based on the lexicon approach. The dictionary created for English Punjabi code mixed consists of opinionated words. The opinionated words are then categorized into three categories i.e. positive words list, negative words list, and neutral words list. The rest of the words are being stored in an unsorted word list. By using the N-gram approach, a statistical technique is applied at sentence level sentiment polarity of the English-Punjabi code-mixed dataset. Our results show an accuracy of 83% with an F-1 measure of 77%.

1 Introduction

In the last decade, the social media platform has been become the medium of communication such as Facebook, Twitter, LinkedIn, etc. (Yang, Chao et al., 2013; Fazil, Mohd et al., 2018). On social media platform, everybody has a short time and the information to be analyzed is huge. Sentiment analysis helps us to whether the message or sentence follows positive or negative. Sentiment analysis is also known as opinion mining or opinion analysis. By using, the web forums there are so many sources to express their views to track and analyze opinions and attitudes about and product. In India, there are 22 official languages, and many more regions languages used for communication (W. Medhat et al., 2014).

There are lots of social media communication, which people use more than one languages to convey their opinion or sentiments (Kalpana et al., 2014; Sharma, S et al., 2015). So, necessary to analyze the data to find appropriate sentiments. N-grams are one of the most commonly used features (G. Rodrigues Barbosa et al., 2012; Kaur, A., & Gupta, V. 2014). We used the n-gram approach up to fivegram and found that the results of fivegram are similar to trigram approach for English-Punjabi code mixed text. The type of ngram also depends on the type of domain used as some domains are more popular in phrases to express the sentiment. Accordingly, our tool gives the power to the users to choose one of two approaches: trigrams and fivegram.

2 Methodology

The main target of current research is sentiment analysis of English-Punjabi code mixed language at sentence level. The foremost task for developing the system is collection of Social Media Code-Mixed text using API twitter threads for **Twitter**, selecting some prolific users comments for **Facebook** as data and some student community prolific users chat for **Whatsapp** followed by cleaning of extracted data.

The dataset used in the current research consists of 10 Lakh sentences (tafter preprocessing) which have been tagged as en

Proceedings of the 17th International Conference on Natural Language Processing: System Demonstrations, pages 24–25
Patna, India, December 18 - 21, 2020.

(English), pb (Punjabi), univ (Universal) and both (mixing of two languages inside a word), The features used are contextual features, capitalization features, special character features, character N Gram features and lexicon features.

In social media text people use creativity in spellings rather than traditional words. The deviation of text can be categorized as acronyms, slangs, misspellings, use of phonetic spellings etc. Contractions like hasn't- has not, ma'am-madam etc. which are handled by mapping. Plenty of common English words e.g. nyt – night, jan- January, gm- gud morning have changed their existence on social media. A dictionary of such out of vocabulary has been maintained in order to normalize them.

3 Results

Generally, the performance of sentiment classification is evaluated by using four indexes: Accuracy with Precision plus Recall and F1-score. A random sample of 200 sentences is picked up for testing and firstly manual testing identified and then tested by a statistical tool. This comparison also discusses the challenges and solutions. We faced and devised on evaluating sentiment analysis. Table 1 represents an accuracy of 83 % with F1-score 77 % on the English-Punjabi code mixed data set the statistical approach.

The accuracy represents the rate at which the method predicts results correctly. The precision also called the positive predictive rate, calculates how close the measured values are to each other. A F- measures that combines precision and recall is the harmonic mean of precision and recall. This score takes both false positives and false negatives into account.

Metrics	**Fivegram Approach**	**Trigram Approach**
Accuracy	82%	83%
Precision	0.83	0.83
Recall	0.71	0.71
F1-Score	0.76	0.77

Table 1: Accuracy, Precision, Recall and F1-score

In order to compute the accuracy of each technique, by calculating the intersections of the positive or negative proportion given by each technique. Table 1 presents the percentage of accuracy for fivegram approach and trigram approach.

References

Fazil, Mohd,and Muhammad Abulaish. A hybrid approah for detecting automated spammers in twitter. IEEE Transactions on Information Forensics and Security, vol. 13, pages 2707-2719, 2018.

G. Rodrigues Barbosa, I. Silva, M. Zaki, W. Meira, R. Prates and A. Veloso, Characterizing the effectiveness of twitter hashtags to detect and track online population sentiment, Proceedings of the 2012 ACM annual conference extended abstracts on Human Factors in Computing Systems Extended Abstracts- CHIEA, vol.1, pages 1186-1195, 2012.

Gelman, A. & Hill, J. Data analysis using regression and multilevel/ hierarchical models, vol.1, pages 1-6, 2007.

Kalpana, R., Shanthi, N., & Arumugam, S. A survey on data mining techniques in agriculture, International Journal of Advances in Computer Science and Technology, vol. 3, pages. 426-431, 2017.

Kaur, A., & Gupta, V. Proposed algorithm of sentiment analysis for punjabi text.Journal of Emerging Technologies in Web Intelligence, vol. 6, pages 180-183, 2014.

Kaur, H., Mangat, V., & Krail, N. Dictionary based sentiment analysis of hinglishtext, International Journal of Advanced Research in Computer Science, vol. 8, pages 1-6, 2017.

Sharma, S., Srinivas, P. Y. K. L., & Balabantaray, R. C. Text normalization of code mix and sentiment analysis, In 2015 international conference on advances in computing, communications and informatics, vol.1, pages 1468-1473, 2015.

W. Medhat, A. Hassan, and H. Korashy, Sentiment analysis algorithms and applications: A survey, Ain Shams Eng. J., no.4, vol. 5, pages 1093–1113, 2014, doi: 10.1016/j.asej.2014.04.011.

Yang, Chao, Robert Harkreader, and Guofei Gu. Empirical evaluation and new design for fighting evolving twitter spammers. IEEE Transactions on Information Forensics and Security, vol. 8, pages 1280-1293, 2013.

NLP Tools for Khasi, a low resource language

Medari Janai Tham
Department of Computer Science and Engineering
Assam Don Bosco University, Assam, India
medaritham16@gmail.com

Abstract

Khasi is an Austro Asiatic language spoken by one of the tribes in Meghalaya, and parts of Assam and Bangladesh. The fact that some NLP tools for Khasi are now available online for testing purposes is the culmination of the arduous investment in time and effort. Initially when work for Khasi was initiated, resources for Khasi, such as tagset and annotated corpus or any NLP tools, were nonexistent. As part of the author's ongoing work for her doctoral program, currently, the resources for Khasi that are in place are the BIS (Bureau of Indian Standards) tagset for Khasi, a 90k annotated corpus, and NLP tools such as POS (parts of speech) taggers and shallow parsers. These mentioned tools are highlighted in this demonstration paper.

1 Introduction

Khasi can be categorized as a low resource language from language technology standpoint due to the fact that reported resources available for Khasi such as a 90k annotated corpus is comparatively small in size (Tham 2020b). However, this did not hinder the development of POS (parts of speech) taggers and shallow parsers for Khasi where their performances are at par with performances of other reported taggers and parsers of other Indian languages. These tools are now available for testing online in https://medaritham.pythonanywhere.com, where users can enter a sentence in Khasi and observe and compare the tagging and parsing performances of various techniques employed in the tools.

2 Khasi BIS Tagset and Annotated Corpus

Due to the unavailability of any corpus in Khasi, the required corpus has to be built from scratch which consumes time and effort. The corpus size is 94,651 tokens extracted from across 38 stories of the prose genre. Details on Khasi corpus construction are explicated in (Tham 2018b) and (Tham 2020b). Next the Khasi tagset is formulated according to the BIS (Bureau of Indian Standards) tagset and the total number of tags is 33. Likewise, further details on Khasi tagset formulation can be found in (Tham 2018b). With the availability of these resources, POS taggers for Khasi were designed and a brief description is given in Section 3.

3 POS Taggers for Khasi

Initially, a Hidden Markov Model (HMM) POS tagger for Khasi was developed and tested only on test data comprising of text from a book not seen during training and achieved an accuracy of 95.68% (Tham 2018b). However rigorous testing of the HMM POS tagger was carried out using ten-fold cross validation giving an accuracy of 93.39%. In order to address the tagging errors of the HMM POS tagger, a Hybrid POS tagger for Khasi was developed which reported an accuracy of 95.29% using ten-fold cross validation (Tham 2020b). This was possible due to the integration of conditional random fields (CRF) which allows the incorporation of language features not possible in an HMM POS tagger. The language features included for Khasi are capitalization, prefixes, current word under consideration, previous word, next word, and whether a word begins or ends a sentence. Due to the absence of inflection, prefixes

Proceedings of the 17th International Conference on Natural Language Processing: System Demonstrations, pages 26–27
Patna, India, December 18 - 21, 2020.

are prevalent in Khasi exhibiting derivational morphology. An additional feature included is the previous tag of a word. Both these POS taggers can be observed and compared online in the site mentioned earlier. One distinct difference is the ability of the Hybrid tagger to differentiate proper nouns from common nouns a trait where CRFs excel.

4 Shallow Parsers for Khasi

To train a shallow parser for Khasi, the annotated corpus has to be further tagged with noun and verb chunks using the BIO labelling specified by Ramshaw and Marcus (1995) where each alphabet symbolizes the following:

B-XX: label **B** for a word starting a chunk of type XX.
I-XX: label **I** for a word inside a chunk of type XX.
O: label **O** for a word outside of any chunk.

Shallow parsing for Khasi has been carried out in the lines of Molina and Pla (2002) where they approached shallow parsing as a tagging problem utilizing the standard HMM approach for parsing without changing the training and tagging process. They have put forward a specialized HMM where adjustments have been made in the training corpus while the training and tagging procedure remains intact. This is carried out by embedding only relevant input information into the model and expanding the chunk tags with supplementary details without affecting the learning and tagging process. The parser is the first of its kind for the language, where the training corpus comprises of 24,194 chunks of noun and verb chunks out of a total of 3,983 sentences and 86,087 tokens. The full details of this process is given in (Tham 2018a). This specialized HMM for Khasi gave an F1 measure of 95.51. This rather optimistic performance is due to the fact that it was tested on gold data that was correctly tagged with POS tags. It was not tested on the output of the POS taggers mentioned earlier. Hence, a deep learning approach using bidirectional gated recurrent (BiGRU) unit was likewise developed for shallow parsing Khasi. The performance of this shallow parser on the same gold data gives an F1 measure of 98.91 (Tham 2020a). However, to get a picture on how its performance will be affected by the performance of the Khasi POS tagger, it was tested on the data tagged by the HMM POS tagger for Khasi and gave an F1 measure of 89.91. This result clearly indicates that improving tagging performance will also improve parsing performance because of the dependency of the parser on POS tagged data. Therefore, the development of the Hybrid POS tagger for Khasi is a much needed improvement in this direction.

5 Conclusion

In this demonstration paper the NLP tools for Khasi and their performances have been presented. Hopefully, with these tools in place it will be rewarding to see their applications in various NLP applications for Khasi.

References

Antonio Molina and Ferran Pla. 2002. Shallow Parsing using Specialized HMMs. *Journal of Machine Learning Research. Introduction to Special Issue on Machine Learning Approaches to Shallow Parsing*, 2:595-613.

Lance Ramshaw and Mitch Marcus. 1995. Text Chunking Using Transformation-Based Learning, in *Proceedings of third Workshop on Very Large Corpora*, pages 82–94.

Medari J. Tham. 2020a. Bidirectional Gated Recurrent Unit For Shallow Parsing. *Indian Journal of Computer Science and Engineering (IJCSE)*, 11(5): 517-521, DOI: 10.21817/indjcse/2020/v11i5/201105167

Medari J. Tham. 2020b. A Hybrid POS Tagger for Khasi, an Under Resourced Language**.** *International Journal of Advanced Computer Science and Applications (IJACSA)*, 11(10):333-342, https://dx.doi.org/10.14569/IJACSA.2020.0111042

Medari J. Tham. 2018a. Khasi Shallow Parser. In *Proceedings of 15th International Conference on Natural Language Processing*. ICON2018, pages 43-49.

Medari J. Tham. 2018b. Challenges and Issues in Developing an Annotated Corpus and HMM POS Tagger for Khasi. In *Proceedings of 15th International Conference on Natural Language Processing*. ICON2018, pages 10-19.

A Chatbot in Malayalam using Hybrid Approach

PRAVEEN PRASANNAN[1] **STEPHY JOSEPH**[2] **RAJEEV R R**[3]
Dept. of Language Technology, International Centre for Free and Open Source Software (ICFOSS)
praveenprasannan@icfoss.org stephy@icfoss.org rajeev@icfoss.in

Abstract

Chatbot is defined as one of the most advanced and promising expressions of interaction between humans and machines. They are sometimes called as digital assistants that can analyze human capabilities. There are so many chatbots already developed in English with supporting libraries and packages. But to customize these engines in other languages is a tedious process. Also there are many barriers to train these engines with other morphologically rich languages. A hybrid chatbot employs the concepts of both Artificial Intelligence (AI) and rule based bots, it can handle situations with both the approaches.

1. Introduction

Chatbot is a computer program that converses with humans like a human partner in natural language. Nowadays, chatbots are being used widely in various domains like Tourism, Health, Business, Customer support. The basic architecture of a chatbot is shown in Figure.1. When a user asks a query to a chatbot, the chatbot in return interacts in human natural language with the users. In general chatbots analyze and identify the user's request intent and extract the relevant entities. After analyzing, an appropriate response will be delivered to the user.

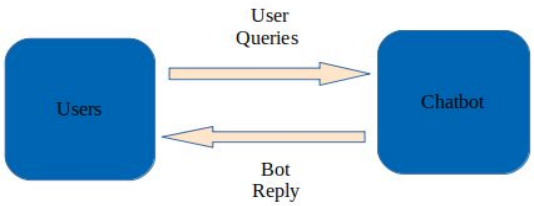

Figure.1

In this work we developed a chatbot in Malayalam in the Pandemic Situation Coronavirus. One of the biggest threats faced by the society during the Corona pandemic was Mis-Information, Dis-information and Mal-information. Government wanted to establish a single source of truth, where the public can rely for authentic information. To support the cause and to fulfill the need to support the general public due to the rapid spread of COVID-19 Pandemic during the months of February and March 2020, we developed an interactive bot which is based on 'hybrid technology' and interacts with the people in regional language (Malayalam).

The main objective of this project was to make the people aware of the disease COVID-19 and the protective measures and precautions to be taken during this period. This chatbot provided all the information related to the pandemic including Government Orders, FAQs and details regarding the COVID-19, authenticated by the State Kerala of Government which is available 24x7 and people do not need to make telephonic contact with the authorities to get the information regarding the virus in this pandemic situation. This chatbot was rolled out through the official website of the State Government of Kerala and was one of the major tools to fight against the situation. The project helped the State Administration as well as the general public in controlling the anxiety among the people. The solution also has the facility of a semi-manual intervention for handling queries with the help of the State Government's Health Department, which the Chatbot is not able to respond to.

2. Methodology

On chatting with the bot, it says, 'കൊറോണ വൈറസുമായി ബന്ധപ്പെട്ടുള്ള വിവരങ്ങൾ അറിയാനായി ചോദിക്കുക', (Ask for information related to the coronavirus). When the user asks about his queries or FAQs,

Proceedings of the 17th International Conference on Natural Language Processing: System Demonstrations, pages 28–29
Patna, India, December 18 - 21, 2020.

the bot assists in autocompleting of his queries or FAQ's. If the user finds any of the autocomplete suitable for his query, he can select that.

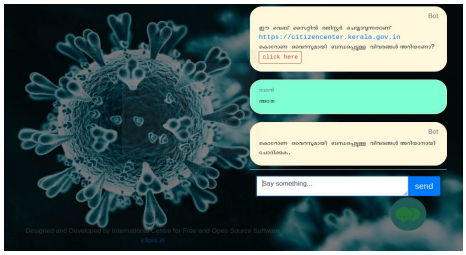

Figure.2

In some cases the user may frame his own questions. The bot should be able to reply in such situations also. But the reply can be relevant or not. In order to make the bot generate a relevant answer, Hybrid approach is used. Hybrid chatbot employs the concepts of both AI and rule based bots and can handle these situations. In this approach, the user query undergoes some preprocessing stages like tokenization, stop words removal and root extraction of keywords. If a match is found, then all the queries regarding that keyword will be the bot's reply and if a match is not found, the system will use similarity measurement to find the most relevant keyword, and respond back.

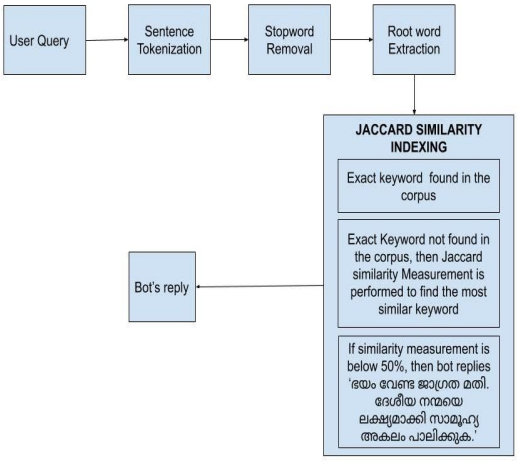

Figure.3

3. Result

When a user asks about the various queries regarding the coronavirus, bot generally categorizes it into three categories (a) exact query available in the database, (b) if exact query not found, then search for the keywords in that query, (c) if exact keyword is not found then use Jaccard indexing. For the above three situations, the chatbot managed to provide relevant responses. For unknown queries a team from Directorate of Health service associated with this work to answer these questions.

4. Conclusion

Already there are so many works going on for the chatbot in English. So there are a lot of libraries and platforms available for building the chatbots in English. Malayalam language is an Indian language that has its own characteristic features, which makes it entirely different from other languages. So there are no readily available modules or platforms for building a chatbot in Malayalam. So developing a chatbot in such a language is a tedious process. Due to the lack of dataset in Malayalam, the accuracy of the chatbot was low in the initial stage. But with the increase in dataset, the accuracy also increased. Thus with the increase in the dataset, the accuracy of the chatbot can be increased.

References

Weizenbaum, Joseph. 1966 "ELIZA—a computer program for the study of natural language communication between man and machine." *Communications of the ACM9.1,* 36-45,, .

Bayan AbuShawar, Eric Atwell. 2015 "ALICE Chatbot: Trials and Outputs" *Computación y Sistemas, Vol. 19, No. 4,,* pp. 625–632 ISSN 2007-9737

Kalaiyarasi, T., Ranjani Parthasarathi, and T. V. Geetha. 2003 "Poongkuzhali-an intelligent tamil chatterbot." *Sixth Tamil Internet 2003 Conference. Vol. 1. sn,.*

S. Chaitrali, Kulkarni, U. Amruta, Bhavsar, Savita Chaitrali S Pingale. 2017. "Bank Chatbot –an Intelligent Assistant System Using Nlp And Machine Learning", *International Research Journal Of Engineering And Technology (Irjet), Volume: 04 Issue: 05, 2017.*

Shah, Rishabh, Siddhant Lahoti, and K. Lavanya. 2017, "Anintelligent chat-bot using natural language processing." *International Journal of Engineering Research6.5 : 281-286. 2017.*

Language Identification and Normalization of Code-Mixed English and Punjabi Text

Neetika Bansal[1], Vishal Goyal[2], Simpel Rani[3]
[1,2]Department of Computer Science, Punjabi University, Patiala, India
[3]Department of Computer Science and Engineering, YCOE, Talwandi Sabo, India
[1]sunshine_neetika@yahoo.com,[2]vishal.pup@gmail.com
[3]simpel_jindal@rediffmail.com

Abstract

Code mixing is prevalent when users use two or more languages while communicating. It becomes more complex when users prefer romanized text to Unicode typing. The automatic processing of social media data has become one of popular areas of interest. Especially since COVID period the involvement of youngsters has attained heights. Walking with the pace our intended software deals with Language Identification and Normalization of English and Punjabi code mixed text. The software designed follows a pipeline which includes data collection, pre-processing, language identification, handling Out of Vocabulary words, normalization and transliteration of English- Punjabi text. After applying five-fold cross validation on the corpus, the accuracy of 96.8% is achieved on a trained dataset of around 80025 tokens. After the prediction of the tags: the slangs, contractions in the user input are normalized to their standard form. In addition, the words with Punjabi as predicted tags are transliterated to Punjabi.

1 Introduction

India is the second largest online market in the world, ranked after China with over 560 million internet users.[1] Facebook is the largest social network with more than 2.7 billion monthly active users followed by WhatsApp, Twitter, and Instagram. Plenty of social media platforms are available nowadays but the most popular in context to Indic languages are Facebook, etc.

(Gold, 1967) was earliest to develop tools for automatic language identification by preparing a Language Learnability Model. (Gumperz, 1962; Scotton, 1997) stated that code-switching occurs when a user switches between different languages in written or spoken a single instance. Nowadays, code switching and code mixing are used alternatively. Word level language identification is one of the challenging tasks as code mixing takes place at word level, at sentence level and even at sub word level in an utterance. Challenges posed are numerous and keep changing with the intensity of languages in the utterance; still due to paucity of data the groundwork remains challenging.

2 Methodology

The main focus of current research is to identify the language of every word in the English Punjabi code mixed. The first and foremost task for developing the system is collection of Code Mixed Social Media Text (English- Punjabi) using API twitter threads for **Twitter**, selecting some prolific users comments for **Facebook** as data and some student community prolific users chat for **Whatsapp** followed by cleaning of extracted data.

(Gamback and Das, 2014) used Hindi, English, acronyms, universal tags along with Code Mixing Index. (Vyas et al., 2014) used English, Hindi and rest tags. In addition to the language tags (Chittaranjan et al., 2014) discussed named entity and ambiguous tags. (Gundapu and Mamidi, 2018)

[1]https://www.statista.com/statistics/262966/number-of-internet-users-in-selected-countries/

Proceedings of the 17th International Conference on Natural Language Processing: System Demonstrations, pages 30–31
Patna, India, December 18 - 21, 2020.

experimented with different possible combinations of available words, context and Part of Speech (POS) tags. (Jamatia et al., 2018) have used Hindi, English, universal, named entity, acronym, mixed and undefined tags.

The dataset used in the current research consists of 80025 tokens (after preprocessing) which have been tagged as en (English), pb (Punjabi), univ (Universal), mixed (mixing of two languages inside a word), ne (Named Entity), acro (Acronyms), rest (none of earlier mentioned tags). A supervised model is trained with Conditional Random Fields (CRF) which calculates the conditional probability of output tags given the values assigned to the input nodes. The features used are contextual features, capitalization features, special character features, character N-Gram features and lexicon features. After applying five-fold cross validation on the corpus, the accuracy of 96.8% is achieved on a trained dataset of around 80025 tokens.

In social media text people use creativity in spellings rather than traditional words. The deviation of text can be categorized as acronyms, slangs, misspellings, use of phonetic spellings etc. Contractions like hasn't- has not, ma'am-madam etc. which are handled by mapping. Plenty of common English words e.g. lyk – like, feb-February, gm- gud morning have changed their existence on social media. A dictionary of such out of vocabulary has been maintained in order to normalize them. A transliterated dictionary for the code mixed data contains transliterated pairs of Romanized text and its Punjabi equivalent. ***e.g*** kithey- ਕਿਥੇ, Janam – ਜਨਮ *etc.*

After the prediction of the tags: the slangs, contractions in the user input are normalized to their standard form words with Punjabi as predicted tags are transliterated to Punjabi language.

3 Results

On the bilingual English-Punjabi data set the CRF baseline approach reports an accuracy of 97.24 % with F1-score 96.8 % on the English-Punjabi language pair. Table 1 shows precision, recall and F1-score with different tag categories used in the system.

Tag Categories	Precision	Recall	F1-Score
acro	0.85	0.77	0.81
en	0.96	0.96	0.96
mixed	0.00	0.00	0.00
ne	0.88	0.92	0.90
pb	0.97	0.99	0.98
rest	0.87	0.54	0.67
univ	0.99	0.94	0.97
Accuracy			0.968

Table 1: CRF System Performance (Accuracy and F1-score) on the Test Data (%)

References

Gokul Chittaranjan, Yogarshi Vyas, Kalika Bali and Monojit Choudhury. 2014. Word-level language identification using CRF: Code-switching shared task report of MSR India system. In *Proceedings of The First Workshop on Computational Approaches to Code Switching,* pages 73-79.

Bjorn Gamback and Amitava Das. 2014. On Measuring the Complexity of Code-Mixing. In *Proceedings of the 11th International Conference on Natural Language Processing*, pages 1-7, Goa.

E. Mark Gold. 1967. Language Identification in the Limit. Information and Control: 10(5), pages 447-474.

John J. Gumperz. 1962. Discourse strategies. Cambridge University Press, Vol.1, Cambridge, UK.

Sunil Gundapu and Radhika Mamidi. 2018. Word Level Language Identification in English Telugu Code Mixed Data. In *PACLIC*.

Anupam Jamatia, Bjorn Gamback, and Amitava Das. 2018. Collecting and annotating Indian social media code-mixed corpora. In *International Conference on Intelligent Text Processing and Computational Linguistics.* pages 406-417, Springer.

Carol Myers-Scotton. 1992. Constructing the Frame in Intrasentential Codeswitching. Multilingua-Journal of Cross-Cultural and Interlanguage Communication, 11(1):101-128.

Yogarshi Vyas, Spandana Gella, Jatin Sharma, Kalika Bali and Monojit Choudhury. 2014. POS tagging of English-Hindi code-mixed social media content. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 974-979.

Punjabi to Urdu Machine Translation System

Nitin Bansal[1], Ajit Kumar[2]
[1] Department of Computer Science, Punjabi University, Patiala, India
[2] Associate Professor, Multani Mal Modi College, Patiala, India
E-mail: [1] profnitinbansal@gmail.com, [2] ajit8671@gmail.com

Abstract

Development of Machine Translation System (MTS) for any language pair is a challenging task for several reasons. Lack of lexical resources for any language is one of the major issues that arise while developing MTS using that language. For example, during the development of Punjabi to Urdu MTS, many issues were recognized while preparing lexical resources for both the languages. Since there is no machine readable dictionary available for Punjabi to Urdu which can be directly used for translation; however various dictionaries are available to explain the meaning of the word. Along with this, handling of OOV (out of vocabulary words), handling of multiple sense Punjabi word in Urdu, identification of proper nouns, identification of collocations in the source sentence i.e. Punjabi sentences in our case, are the issues which we are facing during development of this system. Since MTSs are in great demand from the last one decade and are being widely used in applications such as in case of smart phones. Therefore, development of such a system becomes more demanding and more user friendly. Their usage is mainly in large scale translations, automated translations; act as an instrument to bridge a digital divide.

1 Introduction

Due to the availability of many regional languages in India, machine translation in India has enormous scope. Human and machine translation have their share of challenges. Scientifically and philosophically, machine translation results can be applied to various areas such as artificial intelligence, linguistics, and the philosophy of language. Various approaches are required in machine translation to make communication possible among two languages. These approaches can be rule-based, corpus-based, hybrid or neural-based. Here, hybrid approach is a combination of two approaches i.e. rule-based and corpus-based mainly. The quality of machine translation systems can be measured mainly using Bi-lingual Evaluation Study (BLEU), where it produces a score between 0 and 1.

Among various regional languages in India, we have chosen Punjabi and Urdu for developing Punjabi to Urdu Machine Translation System (PUMTS). Punjabi is the mother tongue of our state, Punjab, where it was used as an official language in government offices. Urdu was also being used as an official language in Punjab, before independence. Thus, PUMTS helps us to make Punjabi understandable to Urdu communities who still want to be in touch with earlier Punjab. These two languages in India, are taken as resource-poor languages, because parallel corpus on language pairs is not available. Thus it became a challenging task for us to develop parallel corpus on this language pair. Further, it also describes types of MTSs being developed with Indian and non-Indian perspective.

2 Methodology

An introduction to Punjabi and Urdu languages help in understanding about history and close proximity among this language pair. Since word-order of this language pair is same but writing order is different from each other i.e. Punjabi can be written from left-to-right and Urdu from right-to-left. Mapping among characters of language pairs has also been studied during the development of PUMTS. The implementation of our methodology for the development of PUMTS, where the architecture followed during the development has been documented. We have proposed three approaches

Proceedings of the 17th International Conference on Natural Language Processing: System Demonstrations, pages 32–34
Patna, India, December 18 - 21, 2020.

to develop bilingual parallel corpus for Punjabi and Urdu languages. But BLEU score suggested for one final approach for corpus development, results in higher accuracy. All the algorithms which were developed during the development of PUMTS, followed the final corpus approach. Lastly, Punjabi to Urdu machine transliteration system to handle Out-of-Vocabulary Words (OOV) words has also been designed and developed, which is working as web-based nowadays. This system has been designed in two phases i.e. first on a web-based platform using ASP.Net and secondly, it has been designed for PUMTS, to handle OOV words during machine translation, using MOSES platform.

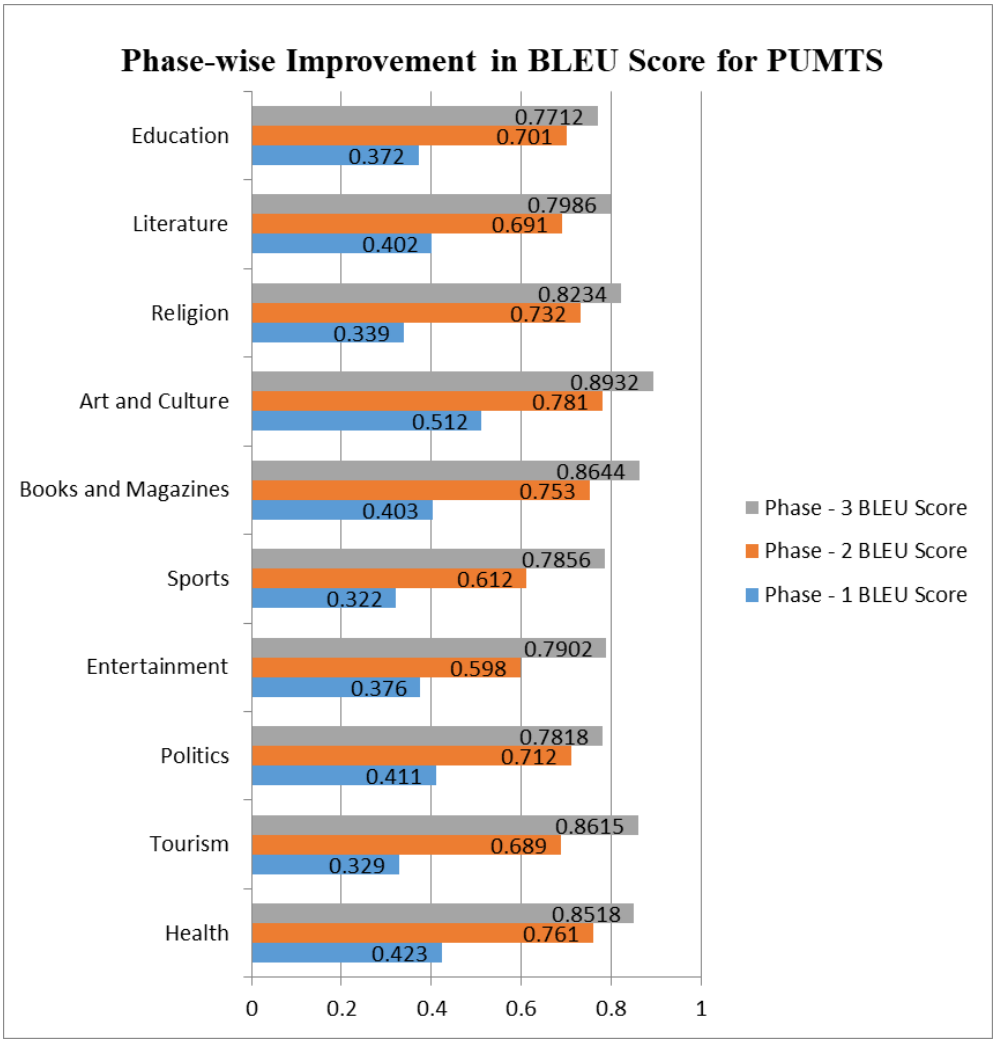

Chart 1: Phase-wise improvement in BLEU score for PUMTS

3 Results and Discussion

Various results had been evaluated by starting from 10000 parallel sentences to 1 lakh parallel sentences after including pre-processing and post-processing modules. The results have been compared with Google translator so as to keep the accuracy comparable and required improvisation can be included in PUMTS. Human evaluation has also been conducted where our evaluators are well known to both the languages. Accuracy has been tested using standard automated metric methodologies i.e. BLEU and NIST, on PUMTS and Google translator. Data domains followed during the development of parallel corpus are politics, sports, health, tourism, entertainment, books &

magazines, education, arts & culture, religion, and literature.

Since, human evaluation is still considered the most reliable and efficient method to test the system's accuracy. However, this is impracticable in today's circumstances. Thus, we have used automatic evaluation with BLEU and NIST to quickly and inexpensively evaluate the impact of new ideas, algorithms, and data sets. During the evaluation of PUMTS, a sufficient bilingual parallel corpus in Punjabi-Urdu language pair (more than 1 lakh parallel sentences) has been used on MOSES, and automated standard metric scores have been generated. Various methods had been applied to increase the system's accuracy, like the order of languages has been changed during the testing to analyze which one gives better results. Moreover, the PUMTS system has also been checked with the Google translator output, where we have found that our system output performs better than Google translator with an accuracy of about 82%. Following chart representation helps us to get an idea where PUMTS generates better results domain-wise.

As shown in chart 1, the development of PUMTS has been started from 10,000 parallel sentences, and the MOSES system has been set-up for this purpose to regularly test the accuracy of this data. Therefore, phase-wise testing and the recording of BLEU and NIST scores has been performed. The second phase has been tested on 50,000 sentences, and after that, final evaluation has been performed on more than 1,00,000 sentences. We can observe from the above chart; there was a sharp increase in accuracy when the number of sentences had been increased from 10,000 to 50,000 sentences. It has also been observed that the increase in size from 50,000 to 1,00,000 results in increments of accuracy at a slower rate, which is due to the handling of OOV words and increments on corpus size, gives more chances of meaningful sentences too.

References

Thomas D. Hedden, 1992-2010, *Machine Translation: A brief Introduction*, http://ice.he.net/~hedden/intro_mt.html

P Koehn, H Huang, et al., 2007, *Moses: Open Source Toolkit for Statistical Machine Translation*. ACL Demos, 2007.

Shahid Aasim Ali and Malik Muhammad Kamran, 2010, *Development of parallel corpus and English to Urdu Statistical Machine Translation*, International Journal of Engineering and Technology, PP. 31-33, Vil 10 No 5, October 2010.

Ajit Kumar and Vishal Goyal, 2011, *Comparative analysis of tools available for developing statistical approach based machine translation system*, in proceedings of International conference ICISIL 2011, Patiala (Punjab), India, PP. 254-260, March9-11.

Tajinder Singh Sani, 2011, *Word Disambiguation in Shahmukhi to Gurmukhi Transliteration*, Processing of the 9th Wordshop on Asian Language Resources, Chiang Mai, Thailand, pages: 79-87, November 12 and 13.

Gurpreet Singh Lehal and Tejinder Singh Saini, 2012, *Development of a Complete Urdu-Hindi Transliteration System*, Proceedings of COLING 2012: Posters, PP. 643-652, COLING 2012, Mumbai.

Arif Tasleem et al, *An analysis of challenge in English and Urdu machine translation*, National conference on Recent Innovations and Advancements in Information Technology (RIAIT 2014), ISBN 978-93-5212-284-4

Ajit Kumar and Vishal Goyal, 2015, *Statistical Post Editing System (SPES) applied to Hindi-Punjabi PB-SMT system*, Indian Journal of Science and Technology", Vol 8(27).

Zakir H. Mohamed and Nagnoor M. Shafeen, 2017, *A brief study of challenges in machine Translation*, International journal of computer Science Issues, PP. 54-57, Vol 14 No 2.

Design and Implementation of Anaphora Resolution in Punjabi Language

Kawaljit Kaur[1], Vishal Goyal[2], Kamlesh Dutta[3]
[1,2]Department of Computer Science, Punjabi University, Patiala, India
[3]Department of Computer Science and Engineering, NIT, Hamirpur, HP,India
[1]saini_kawal@rediffmail.com,[2]vishal.pup@gmail.com,[3]kdnith@gmail.com

Abstract

Natural Language Processing (NLP) is the most attention-grabbing field of artificial intelligence. It focuses on the interaction between humans and computers. Through NLP we can make thec omputers recognize, decode and deduce the meaning ofhuman dialect splendidly. But there are numerous difficulties that are experienced in NLP and, Anaphora is one such issue. Anaphora emerges often in composed writings and oral talk. Anaphora Resolution is the process of finding antecedent of corresponding referent and is required in different applications of NLP.Appreciable works have been accounted for anaphora in English and different languages, but no work has been done in Punjabi Language. Through this paper we are enumerating the introduction of Anaphora Resolution in Punjabi language. The accuracy achieved for the system is 47%.

1 Introduction

Humans have an incredible capability to interact or communicate with one another. Language acts as a powerful medium for this communication. It helps people to express their thoughts using different words but it is quite complex. Complexity of the language can be reflected from the fact that same thing can be narrated in a different no. of ways and a sentence can be interpreted by various people in distinct ways.

A crucial element of the language is the occurrence of the reference which is the process of using language representation to select an "entity" in the real world. The "entity" can be an object in the physical world or a concept in our mind. The linguistic symbol is termed as "referring expression". The issue of reference is the problem of establishing a relationship between different parts of discourse to understand the content which is of great importance for a computational linguistic.

The importance of the problem of reference can be understood from its real time applications in NLP which are - Question Answering/Information Extraction, Automatic Summarization, and Machine Translation as represented in (Aalan F. Smeaten,1994).

ANAPHORA resolution is one of the most sophisticated and complicated problem in modern language processing. The problem of research work is to put focus on Punjabi language in contrast of ANAPHORA RESOLUTION and have fine understanding of the interaction between the syntax and semantics of the language. There are various types of anaphora but pronouns as anaphora are more frequently encountered. So, focus will be on resolving anaphors that act as pronouns.

2 Main Components Of Anaphora Resolution System

The main components of Anaphora Resolution System are:

2.1 Standard pre-processing tools

At the initial stage, the text is required to be converted into the form so that it can be processed by the system. Punjabi Shallow Parser has been used during the pre-processing stage, which performs Morphological analysis and POS tagging. The output is in SSF. Corpus is manually annotated with attributes given in Table-1:

Proceedings of the 17th International Conference on Natural Language Processing: System Demonstrations, pages 35–36
Patna, India, December 18 - 21, 2020.

S.No	Attribute	Used with	Explanation
1	name	All chunks	Gives identification to all phrases within the sentence e.g if it is first NP in the sentence then it will be given name as NP, next NP found in the sentence will be given name= 'NP2' and so on.Similary for other phrases.
2	ref	Pronouns/anaphoras	This establishes anaphoric link.It contains name of the phrase which acts as antecedent for the anaphora.If corresponding antecedent (e.g' NP1')is in same sentence then ref= 'NP1' .But if antecedent is in different sentence then ref='..%2%NP1' i.e. pronoun refers to NP1 of sentence no 2.
3	refType	Pronouns/anaphoras	It specifies the type of anaphora. It can take value 'C' for concrete anaphora or 'E' for event/abstract anaphora.
4	dir	Pronouns/anaphoras	It specifies the direction of antecedent in the discourse. It can take value "A" if antecedent/referent is anaphora(backward direction. It can have value 'C' if antecedent/referent is cataphora (forward direction)
5	PType	Pronoun Type	It specifies type of pronoun: PER3,PER2,PER1 …Third Person/Second Person/First Person REF….Reflexive REL….Relative
6	Animacy	Antecedent	Human,Animate,Inanimate
7	NER	Antecedent	PERSON,ORG,LOC,FAC

Table 1: Annotation Attributes

2.2 Task specific pre-processing module (NP finder and Pronoun Finder)

It deals with extracting Noun Phrase and Pronouns from the sentence. It takes input as sequence of parsedtokens, identifies NP's and discards rest of the tokens. It is used during the step of finding the anaphora and its possible antecedents.

2.3 Feature Extraction Module

Machine Learning Approach has been used as computational strategy and feature extraction is a very important task. It represents data as feature vector of attributes and value pairs. The features describe the properties of anaphora, its antecedent candidates or their relationship. The features that have been extracted from text are- Agreement Feature ,Distance Features, Animacy, NER , Pronoun Type and Referrent Type(Dakwale et al,2012).These features are then applied on model for classification.

2.4 Classification Method

Classification Methods are used to predict most appropriate antecedent for the anaphora. It takes input from Feature Extraction module, performs processing based on set of rules and find the unambiguous antecedent for the anaphora. The classification method which has been employed is Naïve Bayes Classifier.

3 Results

Presently, the system has been checked on a single file having 238 pronouns. The classifier used is Naïve Bayes Classifier. Efficiency of the system is measured using Recall and F score.
Recall:0.46 ,F-Score: 0.63
Overall Accuracy: 47%
Similar results were found during initial work in Hindi language also.

References

Aalan F. Smeaten, Progress in the Application of Natural Language Processing to Information Retrieval Tasks, The Computer Journal, Volume 35, Issue 3, 1992, pp 268-278.

Dakwale, Praveen, Himanshu Sharma, and Dipti Misra Sharma. "Anaphora annotation in hindi dependency treebank." In *Proceedings of the 26th Pacific Asia Conference on Language, Information, and Computation*, 2012, pp. 391-400.

Airport Announcement System for Deaf

Rakesh Kumar[1], Vishal Goyal[2], and Lalit Goyal[3]
[1]Assistant Professor, University College Miranpur, Patiala, rakesh1404@gmail.com
[2]Professor, Punjabi University, Patiala, vishal.pup@gmail.com
[3]Associate Professor, DAV College, Jalandhar, goyal.aqua@gmail.com

Abstract

People belonging to hearing-impaired community feels very uncomfortable while travelling or visiting at airport without the help of human interpreter. Hearing-impaired people are not able to hear any announcements made at airport like which flight heading to which destination. They remain ignorant about the choosing of gate number or counter number without the help of interpreter. Even they cannot find whether flight is on time, delayed or cancelled. The Airport Announcement System for Deaf is a rule-based MT developed. It is the first system developed in the domain of public places to translate all the announcements used at Airport into Indian Sign Language (ISL) synthetic animations. The system is developed using Python and Flask Framework. This Machine Translation system accepts announcements in the form of English text as input and produces Indian Sign Language (ISL) synthetic animations as output.

1 Introduction

Hearing-impaired people use sign language that is regarded as one of the 136 different language families from around 7105 known living languages of the world to convey their feelings and messages (Goyal and Goyal, 2016). Different region of the world have different 136 sign languages worldwide. Nearly 72 million of the nearly 7 billion people on earth are hearing-impaired. Only around 4.3 million sign language users are available from 72 million people. The rest of almost 67 million hearing-impaired people does not use any sign language for communication purpose. So almost 90% of the hearing-impaired people have very limited or no access to education and other information (Goyal and Goyal, 2016). Hearing-impaired people use sign language involving different hand forms, fingers positions, face expressions, hand gestures and other body postures (Goyal and Goyal, 2016). It is a visual-spatial language, as the signer often describes an event using the 3D space around his body (Anuja et al., 2009). As sign languages lack well-defined structure or grammar rules rendering signs less acceptable to outer world of hearing-impaired community. Until the 1960s, sign languages were not taken as bona fide languages but merely grouping or collections of gestures and mime. Dr. Stokoe's American Sign Language research was a great boast to render sign language a full-fledged language with its own grammar rules, syntactic and other linguistic rules/attributes. There are some other contributions to prove the same for other sign languages including the Indian Sign Language (Anuja et al., 2009).

2 Literature Review

The INGIT (Kar et al., 2007) system built for limited domain of railway reservation translates Hindi strings entered as input by reservation clerk to Indian Sign Language (ISL) in gloss string that is then converted using HamNoSys to animated human avatar. The INGIT system works in four modules. ISL generator module applies the ISL grammar rules to get the ISL tagged string. Then each word of the ISL tagged string is replaced with corresponding HamNoSys notation, which are animated using avatars after the conversion into SiGML tags. INGIT is based on hybrid formulaic grammar approach that uses fluid construction grammar (FCG). This project is validated only on a very small corpus of 230 utterances. A

Proceedings of the 17th International Conference on Natural Language Processing: System Demonstrations, pages 37–39
Patna, India, December 18 - 21, 2020.

Prototype Text to ISL is a MT (Dasgupta and Basu, 2008) system based on rules of grammar transfer consisting 5 modules of Pre-processor and parser input text, representation of the LFG f-structure, rules of grammar transfer, generation of ISL sentences , and ISL synthesis. Applying proper transfer rules, the English word structure is translated to an ISL word structure. Then ISL sentences are outputted using a stream of pre-recorded videos or icons. Currently, the system is evaluated based on a set of 208 sentences that gives lexical conversion accuracy of 89.4 per cent. A Frame Based Machine Translation System (Anuja et al., 2009) for English to ISL designed for speech to Indian Sign Language in railways and banking domain. Their system has three modules, speech recognition module that takes input as clerk's speech, translates each uttered phrases into signs language using language-processing module and produces three-dimensional virtual animation using 3D animation module performing signs according to input speech on a display device. Their system use pre-recorded animations. The evaluation of system is done over 250+ phrases with 60% correct translation accuracy, 21% translation with semantic error and 19% incomprehensible translation. An Automatic translation system for English text to ISL (Goyal and Goyal, 2016) consists of parsing, elimination, sentence reordering, lemmatization, words to ISL and animation modules. English words are parsed to Phrase Structure Grammar (PSG) using ISL grammar rules by parsing module. Then eliminator module is used to remove any unwanted words from the sentence reordering module. Then root word is found using lemmatization. The synonym replacement module replaces the non-available word with its synonym words. Then each word is converted into HamNoSys notations that are converted into their respective SiGML tags. Synthetic animation module takes SiGML code as input and generates synthetic animation using various avatars.

3 Methodology

3.1 Creation of Bilingual English-ISL Dictionary

For our research work, we have collected announcements by visiting various airports. Firstly, all airport announcements are broadly categorized into static as well as dynamic announcements that are further sub divided and bifurcated into different categories. Then distinct words are extracted from these announcements. The list of total 1146 distinct words of airport is constructed which are then translated into ISL with help of ISL Teacher and a video footage of all words is prepared. Then these words are coded into HamNoSys one by one to create SiGMLs adding all non-manual components into each sign. Therefore, a bilingual English-ISL dictionary of 1146 distinct airport words is prepared for our system.

3.2 System Architecture

The system will have three modules. Firstly, the system categories the Input text as static, dynamic or randomly generated sentences with the help of corpus of various announcements used at airports.

A) **Mapping Module**: Static and dynamic announcements are passed to Mapping Module that directly maps English words into ISL gloss with the help of bilingual English to ISL dictionary where the static sentences are immediately passed to translation module for generation of synthetic animations. Dynamic sentences are passed to translation module after the replacement of dynamic parts.

B) **Text Processing Module**: Randomly generated sentences are passed to text processing module for parsing the English text using Stanford parser. Then phrase recording and eliminator module is used, to reorder the English sentences and for removal of unnecessary or unwanted word using ISL grammar rules, respectively. Then the root form of English word is obtained applying lemmatization rules after stemming which is passed to translation module for generation of synthetic animations.

C) **Translation Module:** It translates all the ISL words generated from mapping as well as text processing module into HamNoSys notations that are then translated into SiGML file of XML tags. Then a SiGML URL application will be used for converting SiGML into avatars performing animations in Indian Sign Language (ISL).

4 Result and Discussions

The overall accuracy of Airport Announcement System for Deaf system is found to be approximately 83%. In case of simple announcements, the output accuracy is tested to

be 82% and above whereas in case of complex and compound announcements the accuracy is found to be 84%. The system performance can be improved by incorporating word sense disambiguation of more ambiguous words. We have also consulted with ISL interpreters and various ISL experts by showing them the results and the response received was very encouraging and motivating. In various deaf schools, the overall translation accuracy was validated by demonstrating the system.

5 Conclusion and Future Scope

An airport announcement system for Deaf has been presented in this paper. The developed system is the first real-domain announcements translation system for Indian sign language. Our focus was to develop a system that can translate all the necessary information for travelling or visiting at public places into Indian Sign Language (ISL). Our system is able to translate all the announcements used at airports into Indian Sign Language (ISL) synthetic animations so that hearing impaired people can visit airports without any problem. In our system, three-dimensional animated avatars are used because pre-recorded video footage requires large amount of memory space for storage. Animated avatars are easy to upload and download, that's why their processing is fast as compared to recorded video footage and images. In the future, the Bilingual ISL dictionary can be enhanced with adding more words to it. In addition, the developed system can be extended into other public places.

References

Anuja.K, Suryapriya.S, and Sumam Mary Idicula. 2009. Design and development of a frame based MT system for english-to-ISL. *World Congress on Nature and Biologically Inspired Computing, NABIC 2009 - Proceedings*, *pages 1382–1387.* https://doi.org/10.1109/NABIC.2009.5393721

Lalit Goyal and Vishal Goyal. 2016. Automatic Translation of English Text to Indian Sign Language Synthetic Animations. In *13th International Conference on Natural Language Processing*, pages. 144-153.

Purushottam Kar, Madhusudan Reddy, Amitabha Mukerjee and Achla M. Raina. 2007. INGIT: Limited Domain Formulaic Translation from Hindi Strings to Indian Sign Language. In *5th International Conference of Natural Language Processing, 2007, pages. 56-59.*

Tirthankar Dasgupta, Sandipan Dandpat and Anupam Basu. 2008. Prototype Machine Translation System from Text-to-Indian Sign Language," In *Proceedings of the 13th Conference on Intelligent User Interfaces, 2008,* pages. 313-316.

Railway Stations Announcement System for Deaf

Rakesh Kumar[1], Vishal Goyal[2], and Lalit Goyal[3]
[1]Assistant Professor, University College Miranpur, Patiala, rakesh1404@gmail.com
[2]Professor, Punjabi University, Patiala, vishal.pup@gmail.com
[3]Associate Professor, DAV College, Jalandhar, goyal.aqua@gmail.com

Abstract

People belonging to hearing-impaired community feels very uncomfortable while travelling or visiting at Railway Stations without the help of human interpreter. Hearing-impaired people are not able to hear any announcements made at Railway Stations like which train heading to which destination. They remain ignorant about the choosing of platform number or counter number without the help of interpreter. Even they cannot find whether train is on time, delayed or cancelled. The Railway Stations Announcement System for Deaf is a rule-based MT developed. It is the first system developed in the domain of public places to translate all the announcements used at Railway Stations into Indian Sign Language (ISL) synthetic animations. The system is developed using Python and Flask Framework. This Machine Translation system accepts announcements in the form of English text as input and produces Indian Sign Language (ISL) synthetic animations as output

1 Introduction

Hearing-impaired people use sign language that is regarded as one of the 136 different language families from around 7105 known living languages of the world to convey their feelings and messages (Goyal and Goyal, 2016). Different region of the world have different 136 sign languages worldwide. Nearly 72 million of the nearly 7 billion people on earth are hearing-impaired. Only around 4.3 million sign language users are available from 72 million people. The rest of almost 67 million hearing-impaired people does not use any sign language for communication purpose. So almost 90% of the hearing-impaired people have very limited or no access to education and other information (Goyal and Goyal, 2016). Hearing-impaired people use sign language involving different hand forms, fingers positions, face expressions, hand gestures and other body postures (Goyal and Goyal, 2016). It is a visual-spatial language, as the signer often describes an event using the 3D space around his body (Anuja et al., 2009). As sign languages lack well-defined structure or grammar rules rendering signs less acceptable to outer world of hearing-impaired community. Until the 1960s, sign languages were not taken as bona fide languages but merely grouping or collections of gestures and mime. Dr. Stokoe's American Sign Language research was a great boast to render sign language a full-fledged language with its own grammar rules, syntactic and other linguistic rules/attributes. There are some other contributions to prove the same for other sign languages including the Indian Sign Language (Anuja et al., 2009).

2 Literature Review

The INGIT (Kar et al., 2007) system built for limited domain of railway reservation translates Hindi strings entered as input by reservation clerk to Indian Sign Language (ISL) in gloss string that is then converted using HamNoSys to animated human avatar. The INGIT system works in four modules. ISL generator module applies the ISL grammar rules to get the ISL tagged string. Then each word of the ISL tagged string is replaced with corresponding HamNoSys notation, which are animated using avatars after the conversion into SiGML tags. INGIT is based on hybrid formulaic grammar approach that uses fluid construction grammar (FCG). This project is validated only on a very small corpus of 230 utterances. A

Proceedings of the 17th International Conference on Natural Language Processing: System Demonstrations, pages 40–42
Patna, India, December 18 - 21, 2020.

Prototype Text to ISL is a MT (Dasgupta and Basu, 2008) system based on rules of grammar transfer consisting 5 modules of Pre-processor and parser input text, representation of the LFG f-structure, rules of grammar transfer, generation of ISL sentences , and ISL synthesis. Applying proper transfer rules, the English word structure is translated to an ISL word structure. Then ISL sentences are outputted using a stream of pre-recorded videos or icons. Currently, the system is evaluated based on a set of 208 sentences that gives lexical conversion accuracy of 89.4 per cent. A Frame Based Machine Translation System (Anuja et al., 2009) for English to ISL designed for speech to Indian Sign Language in railways and banking domain. Their system has three modules, speech recognition module that takes input as clerk's speech, translates each uttered phrases into signs language using language-processing module and produces three-dimensional virtual animation using 3D animation module performing signs according to input speech on a display device. Their system use pre-recorded animations. The evaluation of system is done over 250+ phrases with 60% correct translation accuracy, 21% translation with semantic error and 19% incomprehensible translation. An Automatic translation system for English text to ISL (Goyal and Goyal, 2016) consists of parsing, elimination, sentence reordering, lemmatization, words to ISL and animation modules. English words are parsed to Phrase Structure Grammar (PSG) using ISL grammar rules by parsing module. Then eliminator module is used to remove any unwanted words from the sentence reordering module. Then root word is found using lemmatization. The synonym replacement module replaces the non-available word with its synonym words. Then each word is converted into HamNoSys notations that are converted into their respective SiGML tags. Synthetic animation module takes SiGML code as input and generates synthetic animation using various avatars.

3 Methodology

3.1 Creation of Bilingual English-ISL Dictionary

For our research work, we have collected announcements by visiting various Railway Stations. Firstly, all Railway announcements are broadly categorized into static as well as dynamic announcements that are further sub divided and bifurcated into different categories. Then distinct words are extracted from these announcements. The list of total 238 distinct words of Railway is constructed which are then translated into ISL with help of ISL Teacher and a video footage of all words is prepared. Then these words are coded into HamNoSys one by one to create SiGMLs adding all non-manual components into each sign. Therefore, a bilingual English-ISL dictionary of 238 distinct Railway words is prepared for our system.

3.2 System Architecture

The system will have three modules. Firstly, the system categories the Input text as static, dynamic or randomly generated sentences with the help of corpus of various announcements used at airports.

A) **Mapping Module**: Static and dynamic announcements are passed to Mapping Module that directly maps English words into ISL gloss with the help of bilingual English to ISL dictionary where the static sentences are immediately passed to translation module for generation of synthetic animations. Dynamic sentences are passed to translation module after the replacement of dynamic parts.

B) **Text Processing Module**: Randomly generated sentences are passed to text processing module for parsing the English text using Stanford parser. Then phrase recording and eliminator module is used, to reorder the English sentences and for removal of unnecessary or unwanted word using ISL grammar rules, respectively. Then the root form of English word is obtained applying lemmatization rules after stemming which is passed to translation module for generation of synthetic animations.

C) **Translation Module:** It translates all the ISL words generated from mapping as well as text processing module into HamNoSys notations that are then translated into SiGML file of XML tags. Then a SiGML URL application will be used for converting SiGML into avatars performing animations in Indian Sign Language (ISL).

4 Result and Discussions

The overall accuracy of Railway Stations Announcement System for Deaf system is found to

be approximately 83%. In case of simple announcements, the output accuracy is tested to be 82% and above whereas in case of complex and compound announcements the accuracy is found to be 84%. The system performance can be improved by incorporating word sense disambiguation of more ambiguous words. We have also consulted with ISL interpreters and various ISL experts by showing them the results and the response received was very encouraging and motivating. In various deaf schools, the overall translation accuracy was validated by demonstrating the system.

5 Conclusion and Future Scope

A Railway Stations announcement system for Deaf has been presented in this paper. The developed system is the first real-domain announcements translation system for Indian sign language. Our focus was to develop a system that can translate all the necessary information for travelling or visiting at public places into Indian Sign Language (ISL). Our system is able to translate all the announcements used at Railway Stations into Indian Sign Language (ISL) synthetic animations so that hearing impaired people can visit Railway Stations without any problem. In our system, three-dimensional animated avatars are used because pre-recorded video footage requires large amount of memory space for storage. Animated avatars are easy to upload and download, that's why their processing is fast as compared to recorded video footage and images. In the future, the Bilingual ISL dictionary can be enhanced with adding more words to it. In addition, the developed system can be extended into other public places.

References

Anuja.K, Suryapriya.S, and Sumam Mary Idicula. 2009. Design and development of a frame based MT system for english-to-ISL. *World Congress on Nature and Biologically Inspired Computing, NABIC 2009 - Proceedings*, *pages 1382–1387.* https://doi.org/10.1109/NABIC.2009.5393721

Lalit Goyal and Vishal Goyal. 2016. Automatic Translation of English Text to Indian Sign Language Synthetic Animations. In *13th International Conference on Natural Language Processing*, pages. 144-153.

Purushottam Kar, Madhusudan Reddy, Amitabha Mukerjee and Achla M. Raina. 2007. INGIT: Limited Domain Formulaic Translation from Hindi Strings to Indian Sign Language. In *5th International Conference of Natural Language Processing, 2007, pages. 56-59.*

Tirthankar Dasgupta, Sandipan Dandpat and Anupam Basu. 2008. Prototype Machine Translation System from Text-to-Indian Sign Language," In *Proceedings of the 13th Conference on Intelligent User Interfaces, 2008,* pages. 313-316.

Automatic Translation of Complex English Sentences to Indian Sign Language Synthetic Video Animations

Deepali[1], Vishal Goyal[2], Lalit Goyal[3]

[1,2]Department of Computer Science, Punjabi University, Patiala, India

[3]Department of Computer Science,Dav College,Jalandhar, India

{singladeepali88,vishal.pup,goyal.aqua}@gmail.com

Abstract

Sign Language is the natural way of expressing thoughts and feelings for the deaf community. Sign language is a diagrammatic and non-verbal language used by the impaired community to communicate their feeling to their lookalike one. Today we live in the era of technological development, owing to which instant communication is quite easy but even then, a lot of work needs to be done in the field of Sign language automation to improve the quality of life among the deaf community. The traditional approaches used for representing the signs are in the form of videos or text that are expensive, time-consuming, and are not easy to use. In this research work, an attempt is made for the conversion of Complex and Compound English sentences to Indian Sign Language (ISL) using synthetic video animations. The translation architecture includes a parsing module that parses the input complex or compound English sentences to their simplified versions by using complex to simple and compound to simple English grammar rules respectively. The simplified sentence is then forwarded to the conversion segment that rearranges the words of the English language into its corresponding ISL using the devised grammar rules. The next segment constitutes the removal of unwanted words or stop words. This segment gets an input sentence generated by ISL grammar rules. Unwanted or unnecessary words are eliminated by this segment. This removal is important because ISL needs only a meaningful sentence rather than unnecessary usage of linking verbs, helping verbs, and so on. After parsing through the eliminator segment, the sentence is sent to the concordance segment. This segment checks each word in the sentence and translates them into their respective lemma. Lemma is the basic requiring node of each word because sign language makes use of basic words irrespective of other languages that make use of gerund, suffixes, three forms of verbs, different kinds of nouns, adjectives, pronouns in their sentence theory. All the words of the sentence are checked in the lexicon which contains the English word with its HamNoSys notation and the words that are not in the lexicon are replaced by their synonym. The words of the sentence are replaced by their counter HamNoSys code. In case the word is not present in the lexicon, the HamNoSys code will be taken for each alphabet of the word in sequence. The HamNoSys code is converted into the SiGML tags (a form of XML tags) and these SiGML tags are then sent to the animation module which converts the SiGML code into the synthetic animation using avatar (computer-generated animation character).

1 Sign Language

Sign language is the mother language for hard of hearing people, it is a physical movement-based language. Sign language is recognized as the first language for the hard of hearing people. In sign language especially hands and head are used to express their thoughts with others.

It is not a universal language; it has its syntax and grammar. Each nation has its sign language. Sign language varies from place to place. Sign language used in the USA is called American Sign Language (ASL), in British they use British Sign Language(BSL). Similarly, in India, we use the Indian Sign Language (ISL).

It is very difficult for the hard of hearing people understand any kind of information because normal people cannot understand their special language and the availability of an interpreter most of the time is not possible. So, these problems make

Proceedings of the 17th International Conference on Natural Language Processing: System Demonstrations, pages 43–45
Patna, India, December 18 - 21, 2020.

it hard of hearing people isolated from the rest of the world.

2 Structure of The English Sentence

The English phrase follows the order of SVO (Subject-Verb-Object). English sentences can be classified into three different forms. These are simple sentences, compound sentences, and complex sentences.

2.1 Simple Sentence

A simple sentence consists of one verb clause. Verb clauses are independent clauses that consist of a subject and a predicate. A simple sentence is a sentence that contain only one independent clause.

Example 1: Boy is singing song.

2.2 Complex Sentence

In complex sentences, an independent clause and dependent clause are joined using subordinate conjunctions. A sentence is called complex sentence that contains at least one dependent clause with independent clause.

Example : If you want to come with us then come on.

2.3 Compound Sentence

A compound phrase is like a pair of twins; each has a different entity, but each has related to the other with the same biological "make-up".

In constructing a compound sentence, there are three key methods: the use of coordinating conjunctions, the use of semicolons, and the use of colons.

The coordinators that are used to join independent clauses are: For, and,Or,But,Nor,Yet.

Example :Both the stream flooded the bridge and a fallen stream blocked the road.

3 Proposed Approaches

The system consists of seven modules:

- English parser is used for parsing the English text
- Sentence reordering module is based on ISL grammar rules.
- Eliminator modules is used for eliminating the unwanted words
- Lemmatization is used for getting the root word of each word to replace the unknown word with its synonyms
- Word to SiGML conversion is done using HamNoSys.
- Synthetic Animation is generated.

4 Results

For evaluating the quality of the English-ISL translator, a set of test sentences is required. The selection of complex and compound sentences for testing the translator has been taken from various English books and internet sources.

A total of 2810 sentences have been tested to evaluate the performance of the translator. To evaluate the translator for English Text to ISL synthetic animations, we have a qualitative system.

In the case of our translator (English Text to ISL Synthetic Animations), the dictionary of ISL is limited to 5370 words including their synonyms. If there is any word in English sentence whose SiGML is not available, then that word is fingerspelled.

Analysis of Sentence Error Rate (SER) Sentence error rate (SER) is the ratio of the number of sentences of MT output which does not match with the reference sentences to the total number of reference sentences.

SER= 471/2810

SER=Number of unmatched sentences/Total number of reference sentences

Out of 2810 sentences, 471 sentences are considered as incorrect, so SER is 0.16.

References

Lalit Goyal, Vishal Goyal "Development of Indian Sign Language dictionary using Synthetic Animations", Indian journal of Science and Technology ,Vol 9(32) (pp 56-59), (August 2016).

Lalit Goyal, Vishal Goyal "Automatic Translation of English text to Indian Sign Language Synthetic Animations", International Conference on Natural Language Processing,(pp. 144-153), (December 2016).

Purushottam Kar, at el. "INGIT: Limited Domain Formulaic Translation from Hindi Strings to Indian SignLanguage",https://www.researchgate.net/publication/228662537,(pp 46-52) (2014).

Kaur, Navneet & Garg, Kamal & Sharma, Sanjeev." Identification and Separation of Complex Sentences from the Punjabi Language". International Journal

of Computer Applications. 13.10.5120/119027976,pp.123-128,2013.

Othman, A., & Jemni, M. (2011). Statistical sign language machine translation: from English written text to american sign language gloss. arXiv preprint arXiv:1112.0168 (pp 56-59).

Dasgupta, T., & Basu, “APrototype machine translation system from text-to-Indian sign language.” In Proceedings of the 13th international conference on Intelligent user interfaces (pp. 313-316). ACM. (2008, January

Plagiarism detection tool for Indian Language documents with Special Focus on Punjabi and Hindi Language

[1]Vishal Goyal, [2]Rajeev Puri, [3]Jitesh Pubreja, [4]Jaswinder Singh
[1,3,4]Punjabi University, Patiala
[2]DAV College, Jalandhar
{vishal.pup,puri.rajeev,jitesh.pubreja,jaswinder.singh794}@gmail.com

Abstract

Plagiarism is closely linked with Intellectual Property Rights and Copyrights laws, both of which have been formed to protect the ownership of the concept. Most of the available tools for detecting plagiarism when tested with sample Punjabi text, failed to recognise the Punjabi text and the ones, which supported Punjabi text, did a simple string comparison for detecting the suspected copy-paste plagiarism, ignoring the other forms of plagiarism such as word switching, synonym replacement and sentence switching etc.

1 Introduction

The above discussed problem led to the scope of development of a specialised software that can bridge the gap. The present software tool aims at providing Successful Recognition and Reporting of Plagiarism among Punjabi and Hindi Documents. This tool is mainly aimed for Universities, Colleges or other Academic Institutions, to check the Plagiarism Report for the Submitted work, be a Ph.D. Thesis, M.Phil. Thesis or a Research Paper of some kind.

The software is built using modular approach and open source technologies. The language dependent components are kept separate from the main programme engine, so that the engine could be used with any other compatible languages. (Figure. 1).

2 Working of Software

1. The Query document is uploaded to the web based interface of the software. The uploaded document can be in plain text form, word document, pdf document or scanned document.
2. The uploaded document is first converted to Unicode format if required. The open source OCR engine Tesseract is used for converting scanned documents to Unicode format.
3. A pre-processing stage performs stop-word removal and stemming of the text, thereby reducing the size of corpus for comparison.
4. The synonym replacement module reduces the document to their base words for comparing with repository and online documents.
5. Keyword identification module identifies the important keywords from the document, that shortlists the documents having higher probability of matches.
6. The similarity matching engine uses the cosine similarity to predict the extent of match of the query document with the online as well as offline sources.

There are Three levels of Users, “Admin”, “Manager” and “User”.

The main Responsibilities for “Admin” are to Maintain and Configure the various aspects of Site for the optimal use of the System. “Admin” can also test the modular systems in isolation to pinpoint the problem if any arises during the Production Environment, which is helpful in Debugging the system down the road without putting it offline.

The “Manager” is the Organizational level head, which can add or manage users to the system under their Domain.

A “User” is the lowest form of Functional account that can be created on this site. Each User can create jobs that contains the documents, articles, papers etc. that needs to be checked.

The Plagiarism Detection tool is under the process of being trademarked in accordance with MietY, under the title “ShodhMapak”. The Copyright has been successfully granted, for the same, to Punjabi University, Patiala in conjunction with MietY.

The software tool is available online for general public at –

Proceedings of the 17th International Conference on Natural Language Processing: System Demonstrations, pages 46–47
Patna, India, December 18 - 21, 2020.

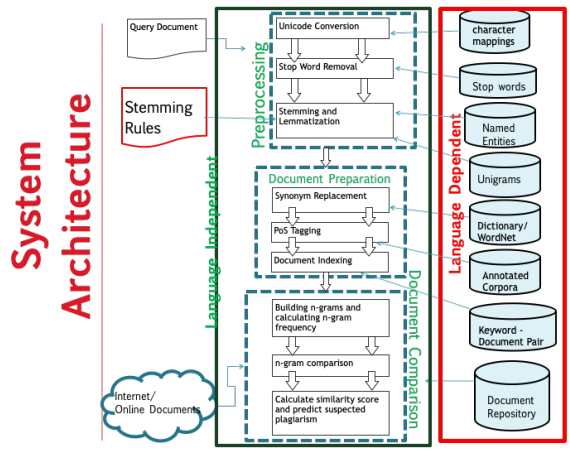

Figure. 1: System Architecture

References

R. Lukashenko, V. Graudina and J. Grundspenk, "Computer-Based Plagiarism Detection Methods and Tools: An Overview," in International Conference on Computer Systems and Technologies, 2007.

B. Martin, "Plagiarism: a misplaced emphasis," Journal of Information Ethics, vol. 3, no. 2, pp. 36-47, 1994.

H. Maurer, F. Kappe and B. Zaka, "Plagiarism - A Survey," Journal of Universal Computer Science, vol. 12, no. 8, pp. 1050-1084, 2006.

Bouville and Mathieu, "Plagiarism: Words and ideas," Journal of Science and Engineering Ethics, vol. 14, no. 3, pp. 311-322, 2008.

R. M. Karp and M. O. Rabin, "Efficient randomized pattern-matching algorithms," IBM Journal of Research and Development, vol. 31, no. 2, pp. 249-260, 1987.

D. Knuth, J. H. Morris and V. Pratt, "Fast Pattern Matching in Strings," SIAM Journal on Computing, vol. 6, no. 2, pp. 323-350, 1977.

R.S.Boyer and J.S.Moore, "A Fast String Searching Algorithm," Comm. ACM. NewYork, NY, USA: Association for Computing Machinery, vol. 20, no. 10, pp. 762-772, 1977.

R. Baeza-Yates and G. Navarro., "A faster algorithm for approximate string matching," Combinatorial Pattern Matching, vol. LNCS 1075, pp. 1-23, 1996.

"Turnitin.com User Guides," iParadigm LLC, [Online]. Available: http://guides.turnitin.com. [Accessed December 2013].

"Urkund,"Urkund,[Online].Available:http://www.urkund.com.[AccessedDecember2013].

"jPlag - Detecting Software Plagiarism," 1996. [Online]. Available: https://jplag.ipd.kit.edu. [Accessed Dec 2013].

L. Bloomfield, "wCopyFind," University of Virginia, [Online]. Available at URL: http://plagiarism.bloomfieldmedia.com /wordpress/ software/wcopyfind/. [Accessed December 2013].

"Dupli Checker - Free Online Software for Plagiarism Detection," Dupli Checker, [Online]. Available: http://www.duplichecker.com. [Accessed 22 10 2015].

W.-Y. Lin, N. Peng, C.-C. Yen and S.-d. Lin, "Online plagiarism detection through exploiting lexical, syntactic, and semantic information," in Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics, Jeju, Republic of Korea, 2012.

S. Schleimer, D. S. Wilkerson and A. Aiken, "Winnowing: local algorithms for document fingerprinting," in SIGMOD, San Diego, 2003.

S. Niezgoda and T.P.Way, "SNITCH-A Software Tool for Detecting Cut and Paste Plagiarism," in 37th SIGCSE technical symposium on Computer science education, 2006.

V. Gupta and G. S. Lehal, "Preprocessing Phase of Punjabi Language Text Summarization," Communications in Computer and Information Science, vol. 139, pp. 250-253, 2011.

V. Gupta and G. S. Lehal, "Features Selection and Weight learning for Punjabi Text Summarization," International Journal of Engineering Trends and Technology, vol. 2, no. 2, pp. 45-48, 2011.

R. Puri, R. P. S. Bedi and V. Goyal, "Plagiarism Detection in Regional Languages – Its challenges in context to Punjabi documents," Research Cell: An International Journal Of Engineering Science, vol. 5, pp. 296-304, 2011.

R. Puri, R. Bedi and V. Goyal, "Automated Stopwords Identification in Punjabi Documents," Research Cell: An International Journal of Engineering Sciences, vol. 8, no. June 2013, pp. 119- 125, 2013.

R. Puri, R. P. S. Bedi and V. Goyal, "Punjabi Stemmer using Punjabi WordNet database," Indian Journal of Science and Technology, vol. 8, no. 27, October 2015.

"Indradhanush WordNet," Dept. of Information technology, Ministry of Communication, Govt. of India.

V. Gupta and G. S. Lehal, "Automatic Keywords Extraction for Punjabi Language," IJCSI International Journal of Computer Science Issues, vol. 8, no. 5, pp. 327-331, 2011.

Author Index