Question and Answer Pair Generation for Telugu Short Stories

Meghana Bommadi, Shreya Terupally, Radhika Mamidi
Language Technologies Research Centre
International Institute of Information Technology Hyderabad, India
meghana.bommadi@research.iiit.ac.in , shreya.reddy@students.iiit.ac.in,
radhika.mamidi@iiit.ac.in

Abstract

Question Answer pair generation is a task that has been worked upon by multiple researchers in many languages. It has been a topic of interest due to its extensive uses in different fields like self assessment, academics, business website FAQs etc. Many experiments were conducted on Question Answering pair generation in English, concentrating on basic Wh-questions with a rule-based approach. We have built the first hybrid machine learning and rule-based solution in Telugu which is efficient for short stories or short passages in children's books. Our work covers the fundamental question forms with the question types: adjective, yes/no, adverb, verb, when, where, whose, quotative, and quantitative(how many/ how much). We constructed rules for question generation using POS tags and UD tags along with linguistic information of the surrounding context of the word.

1 Introduction

Question and Answer pair generation is an open problem in linguistics which deals with Natural Language Understanding (NLU) and Natural Language Generation (NLG). NLU and NLG are commonly used in interactive NLP applications such as AI-based dialogue systems/voice assistants like SIRI, Google Assistant, Alexa, and similar other personal assistants. Numerous methods are introduced for the Q&A pair generation problem. For a low-resourced language like Telugu, AI-based solutions can be non-viable. There are hardly any datasets available for the system to produce significant accuracy. A completely rule-based system might leave out principle parts of the abstract. There is a chance that all the questions cannot be captured inclusively by completely handwritten rules. Hence, we wanted to introduce a mixed rule-based and AI-based solution to this problem.

We attempted to produce questions, concentrating on the key points of a text that are generally asked in assessment tests. Questions posed to an individual challenge their knowledge and understanding of specific topics, so we formed questions in each sentence in as many ways as possible. We based this paper on children's stories, so the questions we wanted to produce aim to be simpler and more objective.

Based on the observation of the data chosen and after analyzing all the possible cases, we developed a set of rules for each part of speech that could be formed into a question word in Telugu. We maximized the possible number of questions in each sentence with all the keywords.

2 Related Work

Previously, Holy Lovenia et al.[2018] experimented on Q&A pair Generation(Holy Lovenia and Gunawan, 2018) in English where they succeeded in forming "What", "Who", and "Where" questions. Rami Reddy et al.[2006] worked on Dialogue based Question Answering System in Telugu for Railway inquiries(Rami Reddy, 2006), which majorly concentrated on Answer Generation for a given Query. Shudipta Sharma et al. worked on implementing automatic Q&A pair generation for English and Bengali texts(Sharma and Hossen, 2018) using NLP tasks like verb decomposition, subject auxiliary inversion for a question tag. Telugu dependency parsing using different statistical parsers (SeshuKumari and RajeshwaraRao, 2017) explored dfferent statistical dependency parsers for parsing Telugu and analysed the performanced of each parser. We explored other[1] Q&A state of art systems from different authors that suita our approach.

[1] (Xu J and R., 2004),(Anne R. Diekema, 2004),(Bert Green, 1961),(Hai and KOSSEIM, 2007)

Proceedings of the 17th International Conference on Natural Language Processing, pages 355–361
Patna, India, December 18 - 21, 2020. ©2020 NLP Association of India (NLPAI)

3 Summarization

Since Telugu is a low resource language, we used statistical and unsupervised methods for this task[2]. Summarization also ensures the portability of our system to other similar low resource languages.

We have used a Telugu stories dataset taken from a website called "kathalu wordpress".[3] This dataset was chosen because of the variety in the themes of the stories, wide vocabulary and sentences of varying lengths. For summarization, we did the basic data preprocessing (spaces, special characters, etc.) in addition to root-word extraction using Shiva Reddy's POS tagger[4].

We implemented two types of existing summarization techniques:

1. Word Frequency-based summarization
2. TextRank based frequency

3.1 Word Frequency-based Summarization

WFBS (Word Frequency-based Summarization)(Shashikanth and Sanghavi, 2019) is calculated using the word frequency in the passage. This process is based on the idea that the keywords or the main words will frequently appear in the text, and those words with lower frequency have a high probability of being less related to the story.

All the sentences that carry major information are produced successfully by this method because the keywords are used repeatedly in children's stories, subsequently causing the highest frequency.

A ratio is used to get a desirable number of sentences in summary (for example: k% of the sentences). If the first highest frequent word is present k out of 100 sentences, we ratio the word selection to 1:n (where n is the total number of words). This ratio, when dynamically changed, performed better than the fixed ratio of word selection.

Steps followed in WFBS are:

1. Sentences are extracted from the input file
2. Words are preprocessed and tokenized
3. Stop words are removed
4. Frequency of each word is calculated
5. The ratio of words that occur in highest to lowest frequency order is calculated

For testing the meticulousness of the user, as a future task, we can use:

1. The least frequent sentences
2. NE (Named Entities) and CN (Common Nouns) to form questions tags (a next level task)

3.2 TextRank based Frequency

TextRank[5] is a graph-based ranking model that prioritizes each element based on the values in the graph. This process is done in following steps:

1. A graph is constructed using each sentence as a node
2. Similarity between two nodes is marked as the edge weight between nodes
3. Each sentence is ranked based on the similarity with the whole text
4. The page-rank algorithm is run until convergence
5. The sentences with Top N ranking as summarized text is given as the output

The TextRank algorithm is a graph based method that updates the sentence score WS iteratively using the following equation(1).

$$WS(V_i) = (1-d) + d * \sum_{V_j \in In(V_i)} \frac{w_{ji}}{\sum_{V_k \in Out(V_j)} w_{jk}} WS(V_j)$$

Where d = damping factor (0.85), w_{ij} is the similarity measure between ith and jth sentences. This method has the advantage of using the similarity between the two sentences to rank them instead of high-frequency words. Two kinds of **similarity** measures used:

Common words: A measure of similarity based on the number of common words in two sentences after removing stop words. We used root word extraction of the common words for better results since Telugu is a fusional and agglutinative language and have repeated words with a different suffix each time.

Best Match 25 : A measure of the similarity between two passages, based on term frequencies in the passage.

The results observed by this method capture the crucial information of the story, but lesser readability and fluency are observed. Within the similarity measures, BM25 has shown slightly better results since the BM25 algorithm ranks based on the importance of particular words (inverse document frequency - IDF) instead of just using the frequency of words.

[2] (Allahyari and Seyedamin Pouriyeh, 2017)

[3] https://kathalu.wordpress.com/

[4] http://sivareddy.in/downloads

[5] (Mihalcea and Tarau, 2004)

4 Answer Phrase Selection

Candidate answers are words/phrases that depict some vital information in a sentence. Adjectives, adverbs, and the subject of a sentence are some examples of such candidates.

The answer selection module utilizes two main NLP components - POS Tagging (Parts Of Speech) and UD parsing (Universal Dependency), along with language-specific rules to determine the answer words in an input sentence.

4.1 POS Tagging

We followed the state of art method called "Cross-Language POS Taggers"(Reddy and Sharoff, 2011) an implementation of a TnT-based Telugu POS Tagger [6] to parse our data.

The tagger learns morphological analysis and POS tags at the same time, and outputs the lemma (root word), POS tag, suffix, gender, number and case marker for each word.

The model was pre-trained on a Telugu corpus containing approximately 3.5 million tokens and had an evaluation accuracy of 90.73% for the main POS Tag.

4.2 UD Tagging

A Bi-LSTM model using Keras is structured and trained using Telugu UD tags dataset "UD_Telugu-MTG". [7]

The Bi-LSTM model outputs the UD Tags for each word in a sentence using Keras. We considered the subject, which is marked "subj" by UD tagger, as a selected answer phrase for a sentence based on a condition that it marked root and punctuation correctly.

This model gave 85% accurate results, including the PAD tags, which might not be an adequate result, but based on the conditions and given that the tags "subj" is labeled in a sentence scarcely, the results have been considered to be acceptable.

4.3 Rules

The outputs of the POS Tagging and UD Parsing modules are used as crucial markers in our language-specific rules. In addition to conditions based on word surroundings, these tags select one or more answer phrases in each sentence.

We classify the rules into different categories, typically based on their usage and interrogative forms.

1. **Quantifiers, Adjectives, Adverbs**: Words with the QC, RB, and JJ POS tag, respectively. For words with JJ tags, the word and corresponding determiners (if present) are selected as the answer candidate.

2. **Possession based**: Words with PRP and NN tags that have suffixes as "టి"," యొక్క", "కి" and "కు" ("ti","yokka","ki" and "ku"). The suffix "టి" ("ti") is used for words like "అతని", "వాళ్ళ", "కంటి", "విద్యార్ధుల" ("athani"-his, "valla"-their's, "kanti"-eyes', "vidyarthula"-students')

3. **Time-Place based** : Noun words with a "లో" ("lo") suffix, along with other words present in custom list of time-related words ("మార్నింగ్","ఇయర్")("morning","year") come under this category.

4. **Direct and Reported Speech**: The word "అని" is generally used to denote direct speech in Telugu. Phrases before the word "అని", along with phrases in quotation marks, are chosen as answer phrases.

5. **Verbs**: Telugu follows the SOV(Subject Object Verb) structure for most of its sentences. If the last word has a "V" POS tag, we selected the verb and adjacent adverbs as an answer candidate.

6. **Subject**: We use the UD tags to determine the subject of a sentence. As an additional check, we only select the candidate subjects in those sentences whose last word is tagged as the root verb, and the subject is a noun.

5 Question Formation

Questions are formed according to the chosen phrases chosen previously, and the question words are replaced using further conditions if required.

1. **Quantifiers, Adjectives, Adverbs**: The words that are marked JJ POS are replaced with "ఎటువంటి" ("etuvanti"- what kind of) RB POS tagged that are followed by verbs with "గా" ("ga") suffix are replaced by "ఎలా" ("ela"-how) and the QC tagged words that are not

[6] https://bitbucket.org/sivareddyg/telugu-part-of-speech-tagger/src/master/

[7] https://github.com/UniversalDependencies/UD_Telugu-MTG, (SeshuKumari and RajeshwaraRao, 2017)

articles ("ఒక" (“oka”- one/once)) were chosen and changed based on the following word. If the quantifier is followed by "శాతం", "మంది" ,"వరకు" (“shatham”,“mandi”,“varaku”) then the word is replaced with "ఎంత" (“entha”-how much), if the quantifier has a suffix it is added to the question word. For example: "1700కు" - "ఎంతకు" (enthaku) and the rest of the quantifiers are replaced with "ఎన్ని" (“enni”-how many).

2. **Possession based**: The Nouns and Pronouns that satisfied the rules are replaced with "ఎవరి" (“evari”-whose) and the dative cases are replaced with "ఎవరికి" (“evariki”-to whom). This could be an exception for non-animus nouns and pronouns. In the children’s stories, most of the nouns are personified, so there were fewer errors than we presumed.

3. **Time-Place based** : We made a list of words that are used to convey time. If the lemma of the word matches the word in the dictionary, then we marked it as “time” and is replaced with "ఎప్పుడు" (“eppudu”-when) or else it is marked as a place and replaced with "ఎక్కడ" (“ekkada”-where).

4. **Direct and Reported Speech** : The whole speech phrase or the phrase that is quoted is replaced with "ఏమని" (“emani”) in the sentence.

5. **Verbs** : The verb is replaced with "ఏమి చేస్తూ" (“emi chesthu”-doing what) + <suffix>”. The appropriate suffix is chosen from the information lost in the lemmatized word.
Additionally, verb tags were used to form polar questions. The interrogative form of a sentence in Telugu can be constructed by adding intonation to the verb, so we added "ఆ" (“aa”) vowel at the end of the verb to make it a yes or no question. The answer phrase to this question would be "అవును" (“avunu”-yes), followed by the original phrase.

6. **Subject** : Based on the suffix of the verb the subject is replaced with "ఏది", "ఏవి" or "దేని", "వేటికి" (meaning what, which simultaneously) or "ఎవరు" (“evaru”-who) and the root suffix is changed accordingly for "ఎవరు" (“evaru”-who).

Question Word	Occurrences	Errors
ఎలా (ela)	64	2
ఎన్ని (enni)	76	5
ఎంతకు (enthaku)	4	0
ఎంత (entha)	3	0
ఎవరి (evari)	187	0
ఎవరికి (evariki)	1	0
ఏమి (emi)	69	3
దేని (deni)	45	10
ఎవరు (evaru)	20	0
ఎప్పుడు(eppudu)	7	0
ఎక్కడ (ekkada)	21	5
ఏమి చేస్తూ (emi chesthu)	148	2
ఏమని (emani)	10	0
ఆ (aa)	148	0
ఎటువంటి (elanti)	103	6
వేటికి (vetiki)	10	1

Table 1: Question Types.

6 Results

We obtained results that resemble commonly used questions covering nine Parts of Speech. The questions generated by this system are successful and are most similar to questions we see in textbooks. In most cases, it has given legible results that resemble human-made questions, with few exceptions for complex sentences. Out of 916 questions formed, only 34 were either completely erroneous or illegible, the rest of them were both grammatically correct and significant for the context of the story.

Table 1 lists the number of times each question word occurred and the number of times it appeared wrong in the experiment with five short stories.

6.1 Error Analysis

Errors are equally influenced by the word tags, the context of the word, and the word’s position in a sentence.

Errors in “ela” (’how’) questions are often caused due to spaces between the words and suffixes in the data set we chose.

“enni” (quantifier - based) questions are built from diverse quantifiers (for example: time, age, number of people - these quantifiers are often written as sandhi with the word, which causes the POS tagger to give ambiguous tags) and numerous ways of writing quantifiers in Telugu. Few quantifier question word errors occurred due to wrong POS tagging of cross-coded words (words that are actually in English but written in Telugu script). In Telugu, two numbers are used together when representing non-specific quantities between the two numbers (x y means from x to y), for example, "rendu(two) moodu(three) nimishalu(minutes)" meaning two to three minutes. This kind of representation makes the system assume there are two quantifies, and the sentence is eligible for two questions based on the same.

"deni" (subject-based) questions have errors because of ambiguous suffixes and inaccuracies in UD tagging. The lack of human identification in the system made human subjects also replaceable with "denini" instead of "evarini". Another error was due to subjects that are names with end syllables similar to common suffixes (which are included as word context in the rule formation). This kind of names were split and formed incorrect question words. The rest of the errors are due to wrong POS tags, cross-codes, and initials/abbreviations.

"emi" ('what') question forms also have similar POS tags and cross-codes issues. Few of these errors occurred due to punctuation marks between the same sentence breaking it up into multiple sentences.

"etuvanti" ('what-kind-of') question forms run into issues where there is personification. General questions based on adjectives for humans are based on a person's subtle qualities; however, in a few cases, the adjective that was chosen is inapt to be formed into a question (less similar to human made question). The question that was formed still is grammatically correct in both human and non-human subjects.

"ekkada" (where) based question forms show errors when an abstract word is used as a place, for example - "In thoughts", "In that age". Certain quantitative words in Telugu can be appended with lo - to convey meanings like "in youth", "in hundreds". They tend to pass the rules in question generation. Our list of time-related words is not exhaustive, so a few time-related words are also tagged under "ekkada" (place) because of the same suffix.

Most of the tags are error-free except for a few ambiguous errors since the rules select answer phrases precisely or do not consider it.

Some of the examples of the questions that are produced by the system are listed below in Table-2 in the appendix.

The results could be improved to make the question formation precise by increasing the number of rules by observing further data.

The anaphora resolution is a limitation in this system; thus, most of the in-appropriation in the answer section was caused due to this.

For example:
Q: ఎవరి చదువంతా సిటీలో, దర్జాగా ... సాగింది?
Q: Whose studies got completed in the city luxuriously?
A: నీ చదువంతా సిటీలో, దర్జాగా ... సాగింది .
A: Your studies got completed in the city luxuriously.

In this case the question is aptly formed but the answer is slightly ill-formed.

There were few errors due to the POS tagger we used. It marked wrong POS tags for cross coded text becuase of the cross coding and the script differences.

7 Conclusions

We have built a mixed rule-based and AI-based question and answer generating system with 96.28% accuracy.We used two methods for summarization and two similarity measures. We constructed observational-based rules for the data set in a particular domain. There is a chance of varying results if we test this system for data in a different domain, but it gives accuracies above 95% for any data in the domain we chose.

We tested question generation in the news article domain, which gave grammatically correct questions. The error rate may increase if we use complex words and phrases that need tags beyond the proposed set of rules.

We plan to extend our work to be able to include:
1.Anaphora Resolution
2.Extending to other domains
3.Cover more types of questions
4.Increase the accuracy of identifying subject for UD tags

References

Mehdi Allahyari and Saeid Safaei Elizabeth D. Trippe Juan B. Gutierrez Krys Kochut Seyedamin Pouriyeh, Mehdi Assefi. 2017. Text summarization techniques: A brief survey.

Elizabeth D. Liddy Anne R. Diekema, Ozgur Yilmazel. 2004. Evaluation of restricted domain question-answering systems. Proceedings of the Conference on Question Answering in Restricted Domains.

Carol Chomsky Kenneth Laughery Bert Green, Alice Wolf. 1961. An automatic question-answerer.

DOAN-NGUYEN Hai and Leila KOSSEIM. 2007. The problem of precision in restricted-domain question-answering. some proposed methods of improvement.

Felix Limanta Holy Lovenia and Agus Gunawan. 2018. Automatic question-answer pairs generation from text.

Rada Mihalcea and Paul Tarau. 2004. Textrank: Bringing order into text. Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing.

Sivaji Bandyopadhyay Rami Reddy, Nandi Reddy. 2006. Dialogue based question answering system in telugu. Proceedings of the Workshop on Multilingual Question Answering - MLQA '06.

Siva Reddy and Serge Sharoff. 2011. Cross language pos taggers for indian languages.

B.Venkata SeshuKumari and Ramisetty RajeshwaraRao. 2017. Telugu dependency parsing using different statistical parsers. *Journal of King Saud University - Computer and Information Sciences*, 29(1):134–140.

Shudipta Sharma and Muhammad Kamal Hossen. 2018. Automatic question and answer generation from bengali and english texts. *Computer Science and Telecommunications 2018*, Volume-54, Issue-2.

Sana Shashikanth and Sriram Sanghavi. 2019. Text summarization techniques survey on telugu and foreign languages. *International Journal of Research in Engineering, Science and Management*, Volume-2, Issue-1.

Licuanan A Xu J and Weischedel R. 2004. Evaluation of an extraction-based approach to answering definitional questions. page 418–424.

A Appendix

Q: ఎటువంటి మోట తో వంగడం కష్టంగా వుంది?
A: అంత పెద్ద మోట తో వంగడం కష్టంగా వుంది

Q: చెప్పు్ప , బట్లు , గాులు , పళ్ళు , గిన్నెలు బజారులో ఎలా కొని , ఊళ్ళో ఇంటింటికి వెళ్లి అమ్ముకునే వాడు?
A: చెప్పులు , బట్టలు , గాజులు , పళ్ళు , గిన్నెలు బజారులో చవకగా కొని , ఊళ్ళో ఇంటింటికి వెళ్లి అమ్ముకునే వాడు

Q: సామాన్లన్నీ మోట కట్టి , గాడిద మీద వేసి , బజారు నుంచి ఊళ్ళో , ఊళ్లోనుంచి తిరిగి ఎవరి ఇంటికి తిప్పేవాడు?
A: సామాన్లన్నీ మోట కట్టి , గాడిద మీద వేసి , బజారు నుంచి ఊళ్ళో , ఊళ్లోనుంచి తిరిగి అతని ఇంటికి తిప్పేవాడు

Q: అమాయక పిచుక ఎక్కడకి, ఎందుకు అని అడగకుండా, ఆ కాకులను గుడ్డిగా నమ్మి ఏమి చేసింది?
A: అమాయక పిచుక ఎక్కడకి, ఎందుకు అని అడగకుండా, ఆ కాకులను గుడ్డిగా నమ్మి వాటితో వెళ్ళింది.

Q: పిచుక మాట నమ్మలేదు కదా , దాని వైపు అసహ్యంగా చూసి మరో ఎన్ని దెబ్బలు వేసారు?
A: పిచుక మాట నమ్మలేదు కదా , దాని వైపు అసహ్యంగా చూసి మరో రెండు దెబ్బలు వేసారు

Q: ఆ కాకులతో పిచుకకి స్నేహం అయ్యిందా?
A: అవును, ఆ కాకులతో పిచుకకి స్నేహం అయ్యింది.

Q: ఒకానొకప్పుడు ఎక్కడ ఒక అమాయకపు పిచుక వుండేది?
A:ఒకానొకప్పుడు ఒక ఊరిలో ఒక అమాయకపు పిచుక వుండేది.

Q: ఏమని పిచుక ప్రాధేయ పడింది?
A:బాబోయ్! బాబోయ్! నా తప్పేమీ లేదు, నేను అమాయకురాలిని, నేనేమీ చేయలేదు, నన్ను వదిలేయండి! అని పిచుక ప్రాధేయ పడింది.

Table 2: Sample questions generated by the system

List of words related to time:
'అప్పుడు', 'రోజు' , 'కాలం', 'సాయంకాలం', 'ఉదయం', 'మధ్యాహ్నం','రాత్రి','పగలు','నెల','వారం','సంవత్సరం', 'సూర్యాస్తమయం', 'శుభోదయం', 'దినం', 'సమయం', 'వర్తమానం' , 'పూర్వం', 'భవిష్యత్తు', 'సోమవారం', 'మంగళవారం', 'బుధవారం', 'గురువారం', 'శుక్రవారం', 'శనివారం','ఆదివారం','మాసం'

Translations Then, day, time period, evening, morning, afternoon, night, morning(synonym), month, week, year, sunset, sunrise, day(synonym), time, present, past, future, Monday, Tuesday, Wednesday, Thursday, Friday, Saturday, Sunday, month(synonym).

This set comprises of the time-related words that have a high chance of being used in a storybook.

Table 3: Time Related word list

Q:What kind of sack was hard to carry?
A:That much of a heavy sack was hard to carry.

Q:In the market how was he buying sandals, clothes, bangles, fruits, utensils - and sold them in the village?
A:In the market how was buying sandals, clothes, bangles, fruits, utensils for cheap rates and sold them in the village.

Q:Packing all the things, putting them on the donkey, from market to village, from village to whose house was he taking them?
A:Packing all the things, putting them on the donkey, from market to village, from village to his own house was he taking them.

Q:How did the innocent sparrow believed the crows without even asking why and where?
A:The innocent sparrow believed the crows blindly without even asking why and where.

Q:Instead of believing the sparrow, looking at it with disgust how many times did they beat it?
A:Instead of believing the sparrow, looking at it with disgust they beat it 2 times.

Q:Did the sparrow made friends with the crows?
A:Yes, the sparrow made friends with the crows.

Q:Once upon a time where was the innocent sparrow living?
A:Once upon a time the innocent sparrow was living in a village.

Q:What did the sparrow say pleadingly?
A:The sparrow said ”No! no! i didn’t any mistake, I’m innocent, I did nothing, Please leave me” pleadingly.

Table 4: Translations of the results in Table 2