Proceedings of the WILDRE-6 Workshop @LREC2020, pages 74–79
Marseille, 20 June 2022

Automatic Identification of Explicit Connectives in Malayalam

Kumari Sheeja S, Sobha Lalitha Devi
AU-KBC Research Centre, MIT Campus of Anna University, Chennai, India
{sheeja,sobha}@au-kbc.org

Abstract

This work presents an automatic identification of explicit connectives and its arguments using supervised method, Conditional Random Fields (CRFs). In this work, we focus on the identification of connectives and their arguments in the corpus. We consider explicit connectives and its arguments for the present study. The corpus we have considered has 4,000 sentences from Malayalam documents and manually annotated the corpus for POS, chunk, clause, discourse connectives and its arguments. The corpus thus annotated is used for building the base engine. The percentage of the performance of the system is evaluated based on the precision, recall and F-score and obtained encouraging results. We have analysed the errors generated by the system and used the features obtained from the anlaysis to improve the performance of the system.

Keywords: discourse, machine learning, discourse relations, conditional random fields, malayalam, NLP

1. Introduction

Discourse analysis is concerned with analysing how phrase, clause and sentence level units of text are related to each other within the larger unit of text. Discourse structure in the documents are important in discourse analysis because it makes the text coherent and meaningful. The pre-processing module also contains the Part of Speech (POS), Chunking and NER which are the essential tool for information retrieval and question answering system. Connectives can be inter or intra sentential. Inter sentential connective occupies the initial position of the sentence which is considered as the one of the argument of the connective and the other argument in the previous sentence. The intra sentential connective appears within the sentence and the two arguments are the main and subordinate clause of the sentence. The clause which follows the connective is the second argument and the other clause is the first argument. The process of discourse annotation involves tokenization and tagging of POS, Chunk and NER using various tag sets.We also annotated the corpus in terms of more basic characterization of discourse structure in terms of identifying discourse connectives in the text and annotating their arguments with semantics. Finally we developed a system for the identification of connectives and their arguments using Machine Learning Technique CRF.

2. Related Work

The annotation study for discourse relations in Arabic (Al-Saif and Markert, 2010) used Machine learning algorithms for automatically identifying explicit discourse connectives and its relations in Arabic language. The sense annotation and sense ambiguities of discourse connectives (Miltsakaki et al., 2005) used syntactic features and simple MaxEnt model and identified several features helped in disambiguation of the connectives since, when and while. (Elwell and Baldridge, 2008) improved the system performance using models for specific connectives and the types of connectives and interpolating them with a general model by using maximum entropy rankers. The work on tagging German discourse connectives using English training data and a German– English parallel corpus (Versley, 2010) and an approach to transfer a tagger for English discourse connectives by annotation projection using a freely accessible list of connectives. This system obtained the F-score of 68.7% for the identification of discourse relations. (Webber et al., 2016) Showed the further procedure of more frequent annotation of more than one discourse relation between the same pair of spans. PDTB annotation also mentioned about another possible source of multiple discourse relations holding concurrently in large corpus. A lexicon of English discourse connectives called DiMLex-Eng (Das et al., 2018) compiled from a lexicon of German discourse connectives DiMLex which focused on modifications to the sense classification of discourse relation in Hindi and comparison based on some initial annotations. (Sobha et al., 2014) used the health domain corpus for the purpose of analysing the discourse connectives and its arguments for languages Hindi, Tamil and Malayalam. They have also presented an automatic discourse relation identifier for the languages Hindi, Tamil and Malayalam (Sobha et al., 2017).

(Faiz and Mercer, 2013) considered the problem of identifying explicit discourse connectives in text using applied machine learning. They used syntactic features to build maximum entropy classifiers for better performance. (Patterson and Kehler, 2013) developed a system for predicting the presence of discourse connectives using classification model. It focused on contrast relation which is the type with lowest model accuracy. An unsupervised approach (Marcu & Echihabi, 2002) to recognize discourse relations that hold between arbitrary spans of texts and show that discourse relation classifiers that use very simple features achieve unexpectedly high levels of performance when trained on extremely large datasets. (Rysova, 2018) introduced the constraints and preferences in the use of discourse connectives in the written Czech texts. An automated system for the identification of arguments of discourse connectives (Wellner & Pustejovsky, 2007) used head based representation for identification of arguments and used re-ranking model for modelling the arguments. Chinese discourse structure (Li et al., 2013) used the advantages of the tree structure from RST and connective from PDTB. Chinese Discourse Tree Bank is developed from 500 Xinhua Newswire documents and employed top-down strategy. Turkish Corpus (Zeyerk and Webber, 2008) based on the principles of PDTB and determined the set of explicit discourse connectives and the syntactic classes. Coordinating and subordinating are not classes in Turkish and most of the existing grammars

of Turkish describe clausal adjuncts and adverbs in semantic rather than syntactic terms.

This work proposes an identification of connectives and arguments in the corpus using the machine learning technique CRFs. Malayalam is a morphologically rich Dravidian language spoken in India. It is a highly inflectional and agglutinative language. It has very different writing style where two or three words are joined together. Features used are rich linguistic features such as suffixes of words, POS, Chunk, Clauses, Connectives and its arguments. The focus of the study is to identify the connectives and its arguments in our corpus. Here Section 3 describes the corpus collection and annotation process. Section 4 describes the method used to develop the system. Feature selection is described in Section 5. The results are discussed in Section 6 and the conclusion is presented in section 7.

3. Corpus Collection and Annotation

A corpus from tourism website consists of 4000 sentences and 47,897 tokens. We have developed the system using machine learning technique CRFs. The training corpus consist of 39,046 tokens and and the testing corpus consists of 8851 tokens. We developed the corpus annotated with discourse connective along with binary arguments by following the guidelines of PDTB (Prasad et al., 2008), a large-scale resource of annotated discourse relations and their arguments. The first argument is tagged as<ARG1> and </ARG1>,the second argument as <ARG2> and </ARG2> and the connective as <CON> and </CON>. The tag sets used for connective and argument identification is described in Table 1.

Sl. No.	Main Tags	Labels
1	Connective begin	<CON>
2	Connective end	</CON>
3	Argument1 begin	<ARG1>
4	Argument1 end	</ARG1>
5	Argument2 begin	<ARG2>
6	Argument2 end	</ARG2>

Table 1: Tag sets-connectives and arguments

In our corpus, the connectives that do not occur as free words were considered to be part of arg1 and the other relation will be arg2. As Malayalam is free word order and inflectional, it consists of many connectives are morphemes and these type of connectives occur intra-sentential. The discourse relation in our corpus can be syntactic (a suffix) or lexical. Discourse relations can be within a clause, inter-clausal or inter-sentential. Examples of annotation of connectives and its arguments are given in example 1 and example 2.

Example 1:

[shareera vedhana kurakkaan idakkide vedhana samharikal upayogich**aal<con><cont-cond>]** </arg1>

body pain decrease frequently pain killer use+**if**

[athu arogyathe dosham cheyyum]/</arg2>

that health harm do

(**If** pain killers are used frequently for body pain, it will be harmful to our health)

In Example 1, the first argument and the second argument are marked as <arg1> and <arg2>.The connective is "**aal"** and it is a subordinate connective and if we take the first argument independently, "shareera vedhana kurakkaan idakkide vedhana samharikal upayogich**aal<con>" (**If pain killers are used frequently for body pain)**,** it will be meaningless. So the use of connectives for relating the arguments makes the text more coherent in the corpus.

Example 2:

[aavaSyaththinu uRakkam kiTTAthe varum**pOL]**</arg1>

enough sleep not+ getting

[SarIraththinu kshINam anubhavappeTunnu]</arg2>

Body tiredness experiencing

(When we do not sleep properly, it leads to body tiredness.)

In Example 2, the verb "varum" (will+come) is combined with the connective **'mpoL' (when),** occur intra-sentential and connects the main clause with the adverbial clause. This representation of the connectives and arguments bring coherence in corpus.

The causal relation is characterized by the cause, the effect and a causal marker. The marker indicates the presence of a causal relation. The cause is the event that has an effect and it acts as a reason for another event to happen. At the discourse level, the causal discourse connective connects two discourse units, arg1 and arg2. The clause that follows the connective is arg2 and the other clause is arg1. The arg2 acts as the reason for the event occurred in arg1.

Example 3:

[Ramuvinu nalla Ormashakthiyundu,] /arg1 **athu kond <con>**

Ramu good memory power so

[avanu pareekshayil mikachcha pragadanam nadaththaan kazhiyum]/arg2

He examination well can+perform

(Ramu has good memory power, **so <con>** he can perform well in the examination.)

The connective "so" in the Example 3 shows "result" relation between two units, where the event in second discourse unit is the result of event in first discourse unit. Here, "Ramuvinu nalla Ormashakthiyundu" (He has good memory power) is the cause and "avanu pareekshayil mikachcha pragadanam nadaththaan kazhiyum"(he can perform well in the examination) is the effect of the cause.

4. Method

The method adopted here uses the machine learning technique CRFs. CRFs uses syntactic and semantic features. These features are obtained by analysing the data. The syntactic features include the suffixes for the words, part of speech, chunk and the clause boundaries and the semantic features include the connective markers and arguments of the connectives.

4.1 Syntactic Pre-Processing

Pre-processing of the text is required as syntactic and semantic information are required for any high level analysis. The pre-processing modules impart the above two information to the text so that the text will have the necessary information required for high end analysis. Syntactic pre-processing is the next step in text analysis after the preliminary processing of sentence splitting and tokenizing. We used the following syntactic pre-preprocessing techniques for developing the connectives and its argument identification of the corpus.

POS Tagger: Part of Speech tagger disambiguates the multiple parse given by the morphological analyser, using the context in which a word occurs. It is the process of tagging the word in a text tagged with its corresponding part of speech such as noun, verb, adjective, adverb, etc. The Part of speech tagger is developed using the machine learning technique Conditional Random fields (CRF++). The features used for machine learning is a set of linguistic suffix features along with statistical suffixes and uses a window of 3 words. The tag set used for developing the POS tagger BIS tag set. These tag sets indicate syntactic classification like noun or verb, and sometimes include additional information, with case markers (number, gender etc.) and tense markers.

Noun Chunk: Chunking is the task of grouping grammatically related words into chunks such as noun phrase, verb phrase, adjectival phrase etc. The system is developed using the feature POS tag, word and window of 5 words.

Clause boundary: Clause is the smallest grammatical unit that has a subject and predicate and expresses a proposition. In Malayalam the subordinate clauses are formed using non-finite verbs. Non-finite verbs are verbs which cannot perform action as the root of an independent clause. The subject of a clause can be explicit or implicit as this language has the subject drop phenomena. In this system we identify the following clauses: Main clause (MC), Relative participle clause (RPC), Conditional clause (CONC), Infinitive clause (INFC), Non-finite clause (NFC), Complementizer clause (COMC). The system is a hybrid system using ML(CRFs) and Linguistic rules combined. The CRFs are trained using annotated corpus and uses linguistic features such as suffix, POS and chunk for learning and mark the beginning and end of a clause. The clause boundary identification depends on word, morphological information and chunk. As begin and end boundaries of the clause matches with the chunk boundaries, chunk boundaries are an important feature of clause boundary identification.

4.2 Semantic Pre-Processing

Once the syntactic pre-processing of the text is over, the system will attempt to produce the logical form of the sentence. The semantic pre-processing is required to ascertain the meaning of the sentence. We used the following semantic pre-preprocessing techniques for developing the system for the identification of explicit connectives and its arguments.

4.2.1 Connective identifier

Connectives are grammatical features such as “but”, “whereas”, which connect two discourse units semantically. The discourse units are called arguments of the connectives. Thus connectives connect two arguments to bring in coherence to the discourse. The discourse unit or the arguments can be intra or inter sentential. If it is intra sentential, it connects the clauses with in a sentence and if it is inter sentential then it connects two sentences.

4.2.2 Discourse argument

The assignment of arguments is syntactic in this work. The arguments can be in the same sentence as the connective or can be outside in the immediately preceding sentence. It is also observed that the argument can be a non-adjacent sentence. But the text span follows the minimality-principle. The position of argument start is on the start of the sentence and this may vary depending on the connectivity with the previous sentence. We used the ML technique CRFs for identifying the beginning and end of each argument.

Example 4 :

[naTuvEdanakku pala kAraNangngaL uNT.] </arg1>

Backpain many reasons are+ there

athinAl<con>

therefore

[yathArththa kAraNam kaNTeththi

Real reason to+be+finding+out

chikilsikkukayANu vENTath.]</arg2>

to+do+treatment

(There are many reasons for back pain. **So** treatment should be taken based on the real reason.)

In Example 4, the connective **“athinaal” (therefore),** occurs inter-sentential by connecting two sentences. Connectives occur at the initial position in the second argument. We see that the connectives are explicitly realizing relations between two arguments arg1 and arg2.

5. Feature selection

Feature selection plays an important role in machine Learning and the learning depends on the features and hence the system's performance. A set of linguistic features are used for identification of connectives and its arguments. The features are discussed below.

5.1 Features for Connective Classification

For connective identification, we used lexical and syntactic features such as word, POS, chunk, clause and

their combinations. The connectives which link groups of words together are mostly conjunction and hence POS features for the identification of connectives is important. Chunk feature segments a sentence into sequence of syntactic constituents and hence it helps to identify the boundary of the connectives and arguments. As connective links clauses or sentences, clause beginning and end of the corpus is used as feature for connective identification.

A baseline system is developed for the identification of connectives using word as feature. Features such as word, POS, chunk, combination of word, POS and chunk and clause are used for developing an extended system. The connective identification for baseline system is performed with minimal features. While developing the baseline system, we considered the first word as one of the features and obtained the f-score as 76.04%. Using word and POS features, we obtained the f-score as 83.47%. This improves the f-score by 7.43%. The inclusion of the chunk feature improves the result by 3.87%. Addition of Clause boundary improves the result by 5.74% and obtained the f-score as 93.08% descibed in Table 2.

Corpus	Word	Word+ POS	Word +POS +Chunk	Word+POS+ Chunk +Clause Boundary
Connectives	76.04	83.47	87.34	93.08

Table 2 : Feature-wise f-measure of connectives

5.2 Features for Argument Identification

The arguments of the connective are also clauses, clause tagging also helps in the identification of the argument boundaries arg1 and arg2. Connectives are used as the key feature in identification of argument boundary. The start and end position of the sentence with respect to the connective are also used as the feature for the identification of arguments. In inter sentential relation, arg1 start and end will be the start and end of the previous sentence of the connective word. The start of arg2 will be after the connective and end of arg2 will be the end of the same sentence. In intra sentential relation, arg2 mostly starts immediately after a connective and ends at the end of the sentence. The arg1 start will be the beginning of the sentence and ends at the clause boundary end in case of intra sentential relation.

Example 5:

[Moonnaar pragrithi souandharyaththinu peru kettathaN] /arg1. **Athinaal<con>**

Munnar scenic beauty known for so [ithu Sthalam sandharshikkan aaLukaLe aakarshikkunnu.]/arg2

it place to visit people attract

(Munnar is known for its scenic beauty. **So** it attracts people to visit the place.)

In the example 5, the connective is inter-sentential, then the end of the preceding sentence is arg1 end, beginning of the preceding sentence is arg1 begin and next token of the connective is arg2 begin, last token of the sentence with connective is arg2 end.

6. Results

In this work, we have used supervised machine-learning approach Conditional Random field for automatically identifying discourse connectives and its arguments of the corpus. This section describes the evaluation and performance of each module using precision, recall and F-score. The evaluation and performance of the system is described using precision, recall, and F-score. For connective identification, we obtained the precision of 92.35%, recall of 93.83% and f-score of 93.08%. The precision of arg1 begin, arg1 end, arg2 begin and arg2 end are 77.81%, 80.14%, 85.53% and 79.28% respectively.

Label	Precision (%)	Recall (%)	F-score (%)
<con>	92.35	93.83	93.08
<arg1>	77.81	79.01	78.39
</arg1>	80.14	83.02	81.54
<arg2>	85.53	87.22	86.35
</arg2>	79.28	80.09	79.69

Table 3: Results of connective and argument identification of corpus

The recall of arg1 begin, arg1 end, arg2 begin and arg2 end are 79.01%, 83.02%, 87.2% and 80.09%. We obtained the f-score of arg1 begin, arg1 end, arg2 begin and arg2 end are 78.39%, 81.54%, 86.35% and 79.67%. This is described in Table3.

The errors generated by the system while classifying the connectives are analysed and the type of errors generated is discussed in the section 6.1.

6.1 Error Analysis

- Some connectives cannot be identified by the system due to some conjunctions identified in POSs which are not connectives and it is not considered as connectives during system identification which affects the performance measures of our system.
- Variation of connective position: Other reasons for occurring errors in the corpus are variation of the position, distribution and sharing of connectives, effect of errors from the previous steps. The connectives such as “allengil”(if+not), “undenkil”(if+so) etc. occur inter or intra sentential depends on the formation of sentence and agglutinative level of the sentence.
- Agglutinative Connective: If corpus contains agglutinative words, system couldn’t identify some of the agglutinated connective words which causes error in connective classification. Connectives such as “vannengil” (if+comes), “poyaal” (if+go), “vannappol” (when+did+come) etc. are morpheme connectives where the verbs are found agglutinated with the connectives “engil”, “-aal”, “-mbol”, “ appoL” (if+so, if, when etc.). Here both lexical and morphemes can become the connectives.

Example 6:

[naam vedhana samharikaL

We pain killers

amithamaaupayogich**aal]** /arg1

if+ over+use

[athu aarogyathe dosham cheyyum]/arg2

that health harm will+do

(If we use more pain killers, it will be harmful to our body)

In Example 6, "amithamaaupayogichaal" (if+over+use) is an agglutinated connective word, and the system fails to identify this type of connective during connective classification.

- Multiple sentences with different connective: Sometimes argument may contain multiple sentences bound with different connectives and it is difficult for the system to identify the position of connectives and results in errors. If multiple connectives (intra and inter sentential) occur in the two consecutive sentences, the system correctly tagged the intra-sentential connective and arguments. At the same time, if the inter sentential connective occurs in the next sentence, the system failed to identify the beginning and end of the first argument of inter sentential connective.
- Most of the errors occur in argument identification is variation of the position of arguments, distribution and sharing of arguments and also the effect of errors from the previous steps.
- When the arguments of the connective appear in different sentences, system couldn't identify the argument boundary of the relation.
- The correlative conjunction such as mAthramalla - pakshe(not only –but also), the system generate errors due to the identification of the pair of conjunctions as a single relation. In this situation, the error occurred in the identification of argument boundaries.

Example 7 :

[bhakshaNakramIkaraNam nAm SIlamAkkiyAl
Dieting we practise

athu SarIravaNNam
that body weight

kuRaykkunnathu **mAthramalla**]/arg1.
Reduce not only

[**pakshe** bhAviyil ArOgyaththOTe
But future health

jIvikkukanum vazhiyorukkunnu.]/arg2
also living will+be+ leading

(Dieting in our daily life **not only** reduces our body weight **but also** helps to lead a healthy life in future)

In example 7, **"maathramalla-pakshe" (not only-but also)** is the correlative connective. But the connective **"pakshe" (but)** is dropped in certain cases.

7. Conclusion

In this work, we have used the syntactic features for identifying the connectives and their arguments in our corpus and consider explicit connectives in our corpus to identify the discourse arguments. In identifying connectives and discourse arguments, we use supervised method, Conditional Random Fields (CRFs). We have developed an annotated corpus of POS, chunk, NER, discourse connectives and its arguments of the corpus. We focussed on the identification of explicit connectives and their arguments in the corpus. We have analysed the errors to improve the performance of the system. In future, we can work with other datasets with better features to improve the performance of the system. We can also work with implicit connectives and arguments of our language based on the semantics and the context of the text by providing a word or phrase to express the relation.

8. Bibliographical References

Al-Saif, A. & Markert, K. (2010). The leeds arabic discourse treebank: Annotating discourse connectives for arabic. *In Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC-2010)*, ed. Calzolari, N, Choukri, K, Maegaard, B, Mariani, J, Odijk, J, Piperidis, S, Rosner, M & Tapias, D, Valletta, Malta, pp. 2046-2053.

Miltsakaki, E., Dinesh, N., Prasad,R., Joshi, A. & Webber, B. (2005). Experiments on sense annotations and sense disambiguation of discourse connectives. *In Proceedings of the Fourth Workshop on Treebanks and Linguistic Theories*,Spain,Barcelona.

Elwell, R. & Baldridge, J. (2008). Discourse connective argument identification with connective specific rankers. *In Proceedings of the IEEE International Conference on Semantic Computing*, ed. Kawada, S, Santa Monica, CA, USA, pp. 198-205.

Versley, Y. (2010). Discovery of ambiguous and unambiguous discourse connectives via annotation projection. *In Proceedings of Workshop on Annotation and Exploitation of Parallel Corpora (AEPC)*, University of Tartu, Estonia, vol.10, pp. 83-92.

Webber, B., Prasad, R., Lee, A. & Joshi, A. (2016). A discourse-annotated corpus of conjoined VPs. *In Proceedings 10th Linguistic Annotation Workshop, Association for Computational Linguistics*, Berlin, Germany, pp. 22–31.

Das, D., Scheffler, T. Bourgonje, P. & Stede, M. (2018). Constructing a lexicon of English discourse connectives. *In Proceedings of SIGDIAL 2018 Conference, Association for Computational Linguistics*, Melbourne, Australia,pp. 360–365.

Sobha, L., Lakshmi, S. & Gopalan, S. (2014). Discourse tagging for Indian Languages. *In Proceedings of CICLing 2014, Springer*, Berlin, Heidelberg, vol 8403, pp. 469-480.

Sobha, L., Sindhuja, G. &.Lakshmi, S. (2017). Cross linguistic variations in discourse relations among indian languages. *In Proceedings of the 14th International*

Conference on Natural Language Processing, NLP Association of India, Kolkata, India, pp. 402-407.

Faiz, S. I., & Mercer, R. E, (2013). Identifying explicit discourse connectives in text. *Advances in Artificial Intelligence, Springer*, Berlin, Heidelber, vol 7884. pp.64-76.

Patterson, G. & Kehler, A.(2013). Predicting the presence of discourse connectives. *In Preceedings of conference on Emperical Methods in Natural Language Processing*, Association for Computational Linguistics,Seattle, Washington, USA, pp. 914–923..

Marcu, D. & Echihabi, A. (2002). An unsupervised approach to recognizing discourse relations. *In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, Philadelphia, Pennsylvania, USA, pp. 368-375.

Rysova, M. & Rysova, K. (2018). Primary and secondary discourse connectives: Constraints and preferences. *Journal of Pragmatics,* vol. 130, pp.16-32.

Wellner, B. & Pustejovsky, J. (2007). Automatically identifying the arguments of discourse connectives. *In Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, Association of Computational linguistics, Prague, pp. 92–101.

Li, Y., Feng, W., & Zhou, G. (2013). Elementary discourse unit in chinese discourse structure analysis. Chinese lexical semantics, Lecture Notes in Computer Science, vol. 7717, pp. 186-198.

Prasad, R., Dinesh, N., Lee, A., Miltsakaki, E., Robaldo, L., Joshi, A. & Webber,B. (2008). The penn discourse treebank2.0. *In Proceedings of the Sixth International Conference on Language Resources and Evaluation*.

Zeyerk, D. & Webber, B. (2008). A discourse resource for turkish: annotating discourse connectives in the METU corpus. *In Proceedings of the 6th Workshop on Asian Language Resources*, Hyderabad, India, pp. 65-72.