Computational Linguistics - 2 Project

Summary of Research Papers

Tanvi Kamble (2018114004) and Sumanth Balaji (2018114002)

30 Nov 2019

1. A Generic Anaphora Resolution Engine for Indian Languages

Vijay Sundar Ram R. , Pattabhi RK Rao and Sobha Lalitha Devi
AU-KBC Research Centre, MIT Campus of Anna University, Chennai, India

This paper aims to generate similarity between the anaphora in Indian Languages (Hindi, Tamil and Bengali) so that we can create an algorithm to work independent of the language used. The paper first discusses how Indian language Anaphora is different from other languages. Indian languages are morphologically rich and verb-final languages. These languages have relatively free word order and clausal structures are more fixed order. Tamil Language belongs to the family of Dravidian Languages and has a rich productive suffixation. Plural marker and case markers get affixed to the nouns and tense markers and Person, Number, Gender (PNG) markers affix with verbs.

This paper uses Machine Learning to form an Anaphora Resolver. For Tamil Language it uses the fact that Pronouns have their Gender and Number markers which help us categorize and select the Candidate Noun Phrases as Antecedent. This essentially means that if a pronoun is marked ’m’ (Masculine) then the nouns having masculine gender are chosen as potential antecedents

The following are the features that are checked for when deciding the appropriate antecedents:

1. Positional Features: Check whether the antecedent and the pronoun lie in the same or different sentences. The antecedent can be present in up to 4 prior sentences.
2. Syntactic Features: These include the role of the antecedent in its sentence. If it is that of an agent or object and so on.
3. Verb Suffixes: The show the gender of the NP

It was noted by observation that more errors were introduced in Tamil language when the pronoun ’avar’ which is used in honorific singular 3rd person. This is due to the possibility of ’avar’ referring to both the masculine and feminine genders

2. Enhancing Tamil WordNet with Subcategorizaton Information

By Vijay sundar rami r and sobha lalitha devi
Anna university chennai

Wordnet is an online lexical reference system. This paper aims to enhance the existing

Tamil wordnet by adding sub categorization as a feature. In Tamil wordnet there exists information about nouns, verbs, adjectives, and adverbs which is organized in the notion of a synset.

In sentences, the verb is the nucleus which imposes restrictions on its arguments known as selection restrictions(SR). Sub-categorization features such as but not limited to +-concrete,+-animate etc can be used to provide more information about the nature of the nouns

In general the ontology is a language based ontology where the nodes in the ontology are sub-categorization features of the nouns. This is different from the taxonomy of nature as the hierarchy is made according to the usage of nouns in the language where each node has a list of nouns as entries of that node.

The ontology starts with a root node “entity” under which any noun can be placed. This then leads to further classifications in a tree like manner such as +-living with +living divided into +-animate and so on.

In wordnet, all nouns are classified as either a physical or an abstract entity. But for the purpose of sub-categorization and SR rules of verbs, this paper suggests the creation of another classification- virtual entities. SR rules analyse a category in terms of syntactic features and are developed according to the type of verb and the number of arguments it can take.

The paper analysed around 2600 verbs that were grouped into 184 groups according to the subject and object it takes. From the analysis it was noted that the most commonly used verbs are more oriented towards human activity. For verbs which take very broad sub-categorization rules, there can be some violations at the finer level.

Hence the paper concludes with its findings.

3. Resolution for Pronouns in Tamil Using CRF

Akilandeswari, A and Sobha, Lalitha Devi
AU-KBC Research Centre, MIT Campus of Anna University, Chennai, India

The main goal of this paper is to develop an automatic anaphora resolution system for Tamil. It includes manual analysis of corpus which contains Tamil pronominal aval, avan, atu and its suffixes. Using the analysis a set of features to identify the anaphoric pronouns and its antecedents is formed. Finally ML algorithm is used.

There are very few works done in anaphora resolution with respect to Indian Languages. Some of the works done are VASISTH a rule based system which works with shallow parsing and exploits the rich morphology in Indian languages for identifying the antecedent for anaphors. (Sobha.L, 2007) used salience measure for resolving pronominals in Tamil. (Murthi.K.N, 2007) have looked into the anaphora resolution in Tamil, using Machine Learning technique: Linear Regression and compared it with salience factors.

Examples are taken that show the use of different pronouns in Tamil based on which certain features are marked to train the ML algorithm. The features are classified into six categories such as word, POS, chunk, syntactic information, clause, Named Entity Recognition. The syntactic information consists of word category,gender,number,person.

The detailed features given in train and test file are sentence recency, subject emphasis, object, proper noun, case of noun phrase, case of anaphor, current clause where anaphor occurs, immediate clause and non-immediate clause of anaphor of the sentence, whether a proper noun followed by a pronoun and NE.

The pronominal avan or aval always refers some person, ultimately a proper noun. avan or aval are the pronoun which are always traverse backward from the current sentence and to other previous sentence to found the antecedent.

Hence it is observed that subject or object which is noun or proper noun in the sentence or clause could be the most probable antecedents for the pronoun. From the analysis we can conclude that, in Indian languages, a nominative noun phrase, a possessive noun phrase with a nominative head and a dative noun phrase could be a subject of a sentence. So we have three types of subject nouns and in that, the most common the nominative noun.

4. Syntactic parser for Tamil

By K. Saravanan, Ranjani Parthasarathi, T.V Geetha
Anna university, Chennai

The paper illustrates a parser for the Tamil language which handles simple and complex sentences with clauses. A parser identifies syntactic constituents of a sentence and represents the same using a parse tree. Many parsers exist that handle fixed order nature languages such as english. This paper aims for a parser for a free word order language i.e Tamil

Generally the tamil sentence follows the subject, object and verb pattern but the interchange of subject and object in the order is acceptable.

Free word order nature of tamil is due to it being a morphologically rich language. This leads to case markers indicating thematic cases indicated by case suffixes to the noun as opposed to position or preposition as in English. This also causes auxiliaries indicating tense, aspect and mood to be attached to the main verb. Tamil also requires person, gender and number agreement between subject and verb. These lead to tamil language requires a different approach to parsing.

The parser is provided with morphologically analyzed sentences as input.

In tamil, the two main components of the parse tree structure are noun constituents(NC) and verb constituents(VC). Since the nouns contain the case markers, there is no preposition phase hence allowing noun constituent to be attached directly with the root of the tree

For simple sentences, adjectives and adverbs are attached with their corresponding words irrespective of the presence of adverb anywhere in the sentence prior to its verb. Then the NC and VC are attached to the root of the tree

For complex sentences with noun clause, adjective clause and adverb clause, they are parsed by first converting the complex sentences to simple sentences by grouping the three clauses with their corresponding words to form NC and VC and then parsing the simple sentence

For complex sentences with more than one verb with conjunctions, first, the sentence is resolved into more than one simple sentence by eliminating the conjunctions. Then for cases where the simple sentence does not have a subject the subject of previous sentence is considered

In conclusion the paper presents a parser capable of handling simple and complex sentences with multiple noun, adjective and adverb clauses and also handles conjunctions to a limited extent.

However there exists obstacles and improvements that need to be tackled with respect to this paper and parser. Subject identification requires gender, animate and inanimate information of nouns while checking the noun and verb coordination. Improvements such as addition of rules for semantic dependencies and ability to produce more than one parse tree for syntactically ambiguous sentences are required. There is also the need to handle syntactically incorrect sentences.