DIALECT RESOLUTION : A HYBRID APPROACH

Sarath K S[*1], Manu.V.Nair[*2], Rajeev R.R[3] and P C Reghu Raj[4]

[*1]Department of CSE, Govt. Engineering College, Palakkad, Kerala, India-678633
sarathks333@gmail.com[1]
[*2]Department of CSE, Govt. Engineering. College, Palakkad, Kerala, India-678633
manunair1990@gmail.com[2]
[3]Research Officer-VRCLC, IIITM-K, Trivandrum, Kerala, India-695581
rajeev@iiitmk.ac.in[3]
[4] Department of CSE, Govt. Engineering. College, Palakkad, Kerala, India-678633
pcreghuraj@gmail.com[4]

Abstract: Dialect Resolution is a regional variety of language, with differences in vocabulary, grammar and pronunciation. It is recognized formal variant of the language spoken by a large group belonging to one region, class or profession. Dialect resolution is an approach to convert a dialect from an informal format to its formal format without losing its meaning. It is a localized approach through which a person can express idea in his own style and still be able to convert to a formal format. In this dialect resolution system, highly informal slang words are replaced by similar words in a given context. This paper presents a dialect resolution method that is a hybrid of Rule based and Machine learning approaches. This work concentrated on Thrissur dialect. Thrissur is a district of Kerala, which is well known for its distinguishable dialect variations.

INTRODUCTION

Language is a social art. It is an exact reflection of the character and growth of its speakers. Variations in language are more often occur in speech than in text. These variations are distributed among a group of people in a geographically separated areas. It is exchanged between the people of the same group for a long period of time and make them adaptive with that style. Cultural, geographical and physical factors also have a role in the art of language. Such a form of a language which is particular to a specific region or social group can be termed as 'Dialect', distinguished by its vocabulary, grammar, and pronunciation.

'Slang' consists of words, expressions and meanings that are informal. They are used either by people who know each other very well or who have the same interests. It include mostly expressions that are not considered appropriate for formal occasions; often vituperative or vulgar. Slang words and phrases highly colloquial and informal in type. It consists either of newly crafted words or of existing words employed in a special sense.

Resolving these dialect and slang words have many applications in day to day life. Localization is the main application of the proposed system in which the local people can engage with outer world, especially for government procedures, easily without language barriers. This system can be embedded with speech to text application systems, that which easily convert the dialect words to formal words. This system can be used in preprocessing stage of Malayalam to other language machine translation systems.

Section 2 describes the common features and characteristics of Thrissur dialect. Section 3 introduces the methods we used in this Dialect Resolution system. Section 4 describes system design and implementation of this model. Section 5 shows observations and experimental result of the proposed system. Section 6 gives the issues related with this system.

RELATED WORKS

There are no works related to Dialect Resolution in Malayalam. Some of the works in dialectal languages are found in Arabic languages. Hassan Sawaf (2010) describe an extension to a hybrid machine translation system for handling dialect Arabic, using a decoding algorithm to normalize nonstandard, spontaneous and dialectal Arabic into Modern Standard Arabic [3]. Wael Salloum and Nizar Habash (2011) proposed a system for improving the quality of Arabic-English statistical machine translation (SMT) on dialectal Arabic text using morphological knowledge [5]. They uses Rule-based approaches in this system.

THRISSUR DIALECT

Thrissur is a cultural city of Kerala. The people keep a unique identity in their dialect and abundant slang words collection, which make them distinguishable in a group immediately. Some slang words are endured and entered the general lexicon and some are used across dialect, like word 'porichu' (means 'fried') used in a sense of 'super' or 'good'. Many metaphor representations are there like 'gadi padayi' with a sense of 'he died' are used frequently. There are many different inflections for dialect and slang words which are easily understandable by people. Through generations, many words get changed resulting in a large dictionary. For example, 'nthu', 'enthutta', 'enthootu', 'nthutta', 'nthootu', 'enthonnu' and 'nthonnu' are

© 2014 Vidya Publications. **Authors are responsible for any plagiarism issues.**

different Thrissur dialects for a single Malayalam word 'enthu' with a meaning 'what'. In a computational point of view, dealing with such properties are computationally intense.

DIALECT RESOLUTION

Dialect resolution system transforms informal dialect sentences into formal, readable, meaning bearing sentences in the same language. In Dialect resolution system, slang words get replaced by its formal equivalents and the dialect word transforms into meaningful words. The system uses rule based method, machine learning method and idea from word sense disambiguation for resolving Thrissur dialect.

Rule based method is typically a mapping concept. In this work, one word is mapped with another word, which is a memorizing task. While dealing with ambiguous words, we need better disambiguating method. Lesk algorithm is a simple and dictionary based approach for word sense disambiguation (WSD) [2]. In this approach, all the sense definitions of the word to be disambiguated are retrieved from the dictionary. Each of these senses is then compared to the dictionary definitions of all the remaining words in the context. The sense with highest overlap with these context words is chosen as the correct sense. The idea of working of Lesk algorithm is used for disambiguation.

Machine Learning

Machine learning is the construction and study of systems that can learn from data, rather than follow only explicitly programmed instructions. It tries to find hidden pattern in the given data and tries to predict the future data. Machine learning is a type of artificial intelligence (AI) that provides computers with the ability to learn without being explicitly programmed [1].

Machine learning is usually divided into two main types, supervised learning and unsupervised learning. In the predictive or supervised learning approach, the goal is to learn a mapping from inputs x to outputs y, given a labelled set of input output pairs. In unsupervised learning set of inputs are modelled without the help of labelled examples.

TnT: Trigrams 'n' Tagger (TnT) is an efficient statistical part-of-speech tagger developed by Thorsten Brants (2000). It is a tagger based on second order Markov model which consider triples of consecutive words, performs at least as well as other current approaches, including Maximum Entropy framework. It works significantly better for the tested corpora. The main specialty is that TnT is independent of language and tagset used. According to tagger evaluation and comparison test, it achieved an average accuracy between 96%–97%, depending on language and tagset [4].

In this system, the change in characters of a dialect word are used as tags for learning in TnT. Transition probabilities depend on the states. Output probabilities only depends on the most recent category. For a given sequence of words $w_1 \ldots w_T$ of length T, calculate

$$\underset{t_1 \ldots t_T}{\operatorname{argmax}} [\prod_{i=1}^{T} P(t_i | t_{i-1}, t_{i-2}) P(w_i | t_i)] \, P(t_{T+1} | t_T) \qquad (1)$$

$t_1 \ldots t_T$ are elements of tagset, the additional tags t_1, t_0, and t_{T+1} are beginning-of-sequence and end-of-sequence markers. Both transition and output probabilities are estimated from a tagged corpus. Linear interpolation is used as smoothing paradigm and the trigram probability is estimated as follows,

$$P(t_3 | t_1, t_2) = \lambda_1 \hat{P}(t_3) + \lambda_2 \hat{P}(t_3 | t_2) + \lambda_3 \hat{P}(t_3 | t_1, t_2) \qquad (2)$$

$\hat{P}$ are maximum likelihood estimates of the probabilities, and sum of three coefficients is 1. Values of coefficients λ_1, λ_2 and λ_3 are estimated by deleted interpolation. The system also take unseen data into account.

SYSTEM DESIGN AND IMPLEMENTATION

The task is a learning processes where each change on formal word is learned through sample large corpus. All possible slang words are substituted, because they gives no information about target word. More context words are required to replace a slang word, sometimes they are not available. More than one dialect word in a sentence is slightly crucial task. Accuracy in the following target words are depends upon the previous resolved formal words. In the same way, treating an unknown word is highly crucial one. Resolved word or words may together deviate from meaning of source sentence. And the words in a sentence are highly dependent.

Design

The design phase is composed of three major levels. Source sentence is passed through the first level (rule-based level) to resolve all the slang words in it. After that, the untouched words are entered into statistical level where they transformed into corresponding formal word as the system learned through corpus. All the resolved known words are together used as context to resolve unknown words one-by-one in the third level (word disambiguation level).

Implementation

In this system, input sentences are treated separately. Sentences with any number of slang are allowed. At most two dialect words per sentence are used in testing phase.

Rule-based level: Slang words are almost fixed, they only varied through generations. Almost all current slang words and their inflections are used for this phase. All slang words in source sentence are replaced by predefined formal word, one after another. This phase is almost error free. Only predefined words are substituted and their possible synonyms are not used.

© 2014 Vidya Publications. **Authors are responsible for any plagiarism issues.**

Machine Learning level: TnT is a statistical machine learning tool used in this level for learning morphological variations on words during transformation of dialect word to formal. TnT uses second order Markov model, hence sequential information in characters are considered. Validity of resolved word is checked using Malayalam formal word dictionary. Sample of training corpora is given in Fig 2. All the valid words are taken as list of context words for the next level.

Word disambiguation level: Concept of word sense disambiguation (Lesk algorithm) is used to handle unknown words. Each unknown word is treated one-by-one from left to right in source sentence. For each word, possible word list is created from formal word dictionary by taking an assumption that first character of unknown and target words are the same. Each word in the list is ranked by a count of number of context words come together in sentences more frequently in formal sentence corpus. The most ranked word is used as target word, and appended to context word- list. Hence following unknown words will get more number of context words. But the total accuracy is inter dependent. Succeeding accuracy is proportional on the preceding resolved words. Order of context words are not considered, because Malayalam is highly free order language.

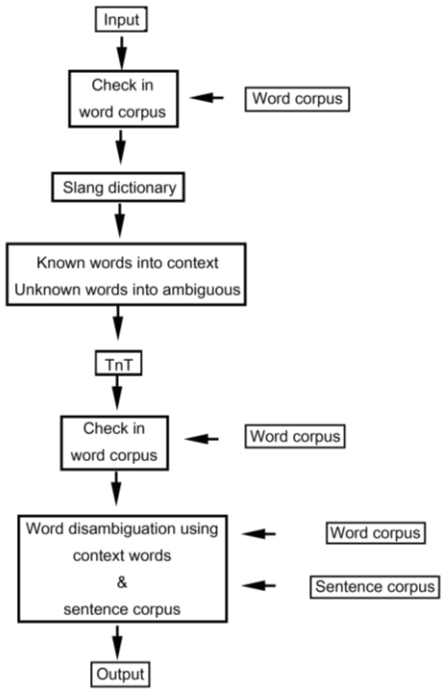

Figure 1. Flow chart of three phases

പ പ
റ റ
യ യു
ണേ ന്നേ
. .
പാ പാ
വാ വമാ
. .
മ മ
റ റി
ച്ചാ ച്ചേ
. .
ഇ ഇ
ര് ര
ട്ട ട്ട
ത്ത് ത്ത്
. .
ഇ ഇ
ര് ര
ന്നാ ന്നോ
ട്ടാ .
. .

Figure 2. TnT Training corpus

TEST AND OBSERVATIONS

For following sample input sentence of six words.

സ്വന്തം കാര്യം നോക്കാണ്ട് കൊറേ പ്രാവും ഇഷ്ടൻ .

Figure 3. Input sentence

Let w_1, w_2, w_3, w_4, w_5 and w_6 are corresponding words in the sentence given in fig. 3. The words w_3, w_4, w_5 and w_6 are unknowns. The word w_6 is a slang and resolved in Rule based step using slang dictionary.

സ്വന്തം കാര്യം നോക്കാണ്ട് കൊറേ പ്രാവും സുഹൃത്ത് .

Figure 4. Slang word correction

The words w_4 and w_5 are resolved by TnT. The tagged output of TnT is given in fig. 5.

```
%% Thorsten Brants, thorsten@brants.net
നോ                  നോ
ക്കാ          ക്കാ
ണ്                  ണ്
ട്                  ട്
.                   .
കൊ                  ക
റേ                  റേ
.                   .
പ്                  പ്
രാ                  രാ
വും           ക്കം
.                   .
%% 0 (0.00%) unknown tokens
%% avg. 1.92 tags/token
```

Figure 5. TnT output

After passing into the TnT, the w_4 and w_5 are resolved.

സ്വന്തം കാര്യം നോക്കാണ്ട്ട് ക്കറേ പ്രാക്കം സുഹൃത്ത് .

Figure 6. After TnT output

***© 2014 Vidya Publications.* Authors are responsible for any plagiarism issues.**

Finally, only w_3 is unknown which is passed to word disambiguation stage and substitute word with highest rank.

അവൾ പാവം ആണ്
സ്വന്തം കാര്യം നോക്കാതെ കൊറെ പ്രാകും
ആ മോശമായവന് നെഗളം കൊറെ കൂടുതലാണ്

Figure 7. Output after word disambiguation level

Rank is assigned for each word in the possible word-list according to the number context words co-occur with it. The maximum ranked word is chosen as target word. The output sentence is given in fig. 8.

സ്വന്തം കാര്യം നോക്കാതെ കൊറേ പ്രാകും സുഹൃത്ത് .

Figure 8. Resolved sentence

The sentence corpus contains 400 sentences. The word dictionary contains 700 formal words. The rule based dictionary contains 160 slang words and their corresponding formal words. Experimental results are shown in Table 1.
According to the results, the proposed system resolved exactly all formal and slang words in the input sentences correctly. The overall accuracy of the system depends on its performance on resolving the dialect words. The system has shown a performance of 61.11% by keeping semantic validity of words in the sentence. There is no previous work done on resolving dialects in Malayalam language.

Table 1. Results

Number of input sentences	***50***
Total number of words in input sentences	190
Average number of words in a sentence	4
Number of sentences correctly resolved	23 (46%)
Total number of slang words correctly resolved	11 out of 11 (100%)
Total number of dialect words correctly resolved	66 out of 108 (61.11%)

ISSUES IN DIALECT RESOLUTION

Dialect resolution in each of its stages have many issues which shows its complexity. This paper is mainly concentrate on Thrissur dialect. Thrissur dialect is complex with compound and slang words. Slang word doesn't give any clue about corresponding formal word from its morphological information. Handling these slang words are very difficult. In machine learning level, named entities may transformed wrongly by TnT and generate morphologically different one.

In this system, at least two strong context words are necessary for one dialect word to be resolved, otherwise prediction will be wrong. Adding wrong prediction to context word-list also reduce contextual information for the following cases in a sentence. Resolving a dialect 'B' after a dialect 'A' is different from resolving 'A' after 'B'. Creating all possible inflectional forms of Malayalam sentences which are used in word disambiguation level, is a tedious task. The sentences having metaphor cannot be included, because of its hidden meaning cannot be correctly mapped.

CONCLUSION AND FUTURE SCOPE

This system is one of the first approach in Malayalam language. This system can also be used for other dialectal languages, with sufficient support of corpus. As an extension of this approach, gender, tense, person and number information can be labelled for disambiguation. It will become better with large corpus of formal inflectional sentences, dictionary with all slang words and large formal word corpus. A better machine learning technique like SVM or CRF with large corpus may provide better result in future.

REFERENCES

[1] T.O Ayodele, "Types of Machine Learning Algorithms", INTECH Open Access Publisher, 2010

[2] D. Jurafsky and J.H Martin, "Speech and Language Processing: an introduction to natural language processing, computational linguistics and speech recognition", Prentice Hall series in artificial intelligence, 1999.

[3] Hassan Sawaf, "Arabic Dialect Handling in Hybrid Machine Translation", in proceedings of the 9th Conference of Association for Machine Translation in the Americas (AMTA), 2010

[4] Thorsten Brants,"TnT-a statistical part-of-speech tagger ", Technical report, Computational Linguistics, Saarland University, 2000.

[5] Wael Salloum and Nizar Habash, "Dialectal to Standard Arabic Paraphrasing to Imprsssove Arabic-English Statistical Machine Translation", in proceedings of the Dialects workshop at the Conference for Empirical Methods in Natural Language Processing (EMNLP 2011), Edinburgh, UK, 2011.

© 2014 Vidya Publications. **Authors are responsible for any plagiarism issues.**