“Nee Intention enti?” Towards Dialog Act Recognition in Code-Mixed Conversations

by

jittadivya.sai , Chandu Khyathi Raghavi, Harsha Pamidipalli, Radhika Mamidi

in

21st International Conference on Asian Language Processing
(*IALP-2017*)

Report No: IIIT/TR/2017/-1

Centre for Language Technologies Research Centre
International Institute of Information Technology
Hyderabad - 500 032, INDIA
December 2017

"Nee Intention enti?" Towards Dialog Act Recognition in Code-Mixed Conversations

Divya Sai Jitta, Khyathi Raghavi Chandu, Harsha Pamidipalli and Radhika Mamidi
International Institute of Information Technology-Hyderabad, India
Email: {jittadivya.sai, chandukhyathi.raghavi}@research.iiit.ac.in
Email: harsha.pamidipalli@students.iiit.ac.in
Email: radhika.mamidi@iiit.ac.in

***Abstract*—Code-Mixing (CM) is a very commonly observed mode of communication in a multilingual configuration. The trends of using this newly emerging language has its effect as a culling option especially in platforms like social media. This becomes particularly important in the context of technology and health, where expressing the upcoming advancements is difficult in native language. Despite the change of such language dynamics, current dialog systems cannot handle a switch between languages across sentences and mixing within a sentence. Everyday conversations are fabricated in this mixed language and analyzing dialog acts in this language is very essential in further advancements of making interaction with personal assistants more natural. The problem is further compounded with crossing the script barriers in code-mixing. In this paper we take the first step towards understanding code-mixing in dialog processing, by recognizing dialog act (intention) of the code-mixed utterance. Considering the dearth of resources in code-mixed languages, we design our current system using only word-level resources such as language identification, transliteration and lexical translation. Our best performing system is HMM based with an F-score of 76.67.**

***Keywords*-Code-mixing;Language-identification;Dialog-acts;Translation;**

I. Introduction

Code-Mixing is defined as the embedding of linguistic units such as phrases, words, and morphemes of one language into an utterance of another language. Apart from being a commonly used spoken form in multilingual settings, CM also manifests itself on social media sites in the form of posts, comments, replies to the comments and most importantly chat conversations. Most informal and semi-formal conversations are fabricated in CM fashion. It usually occurs in scenarios where the formal education is received in a language other than the person's native tongue. Previous studies on the reasons for facebookers to switch language is 45% due to real lexical need, which resulted in 58.97% of inter-sentential switching and 33.33% of intra-sentential switching [1]. Popularly used personal assistants like Siri, Cortana, Alexa etc., currently do not handle this case of language switching in the course of a conversation.

In addition to the use of borrowed words when an equivalent word is absent in the dominant language, mixing happens with very commonly used words as well. For example, consider the following sentence from a website called *chai-bisket*[1], comprising of general articles in cross script code-mixed languages (representative of the flavour used in social conversations and chats). Data scraped from this site is used to develop a language identification module, which would be discussed in the following sections.

An example sentence representative of the mixing being dealt in this paper is as follows: *Anni_Tel subjectlu_Tel,Eng clear_Eng chesaru_Tel kani_Tel subject_Eng knowledge_Eng sunna_Tel. (Translation: They have cleared all the subjects, but their subject knowledge is zero.)*. The words followed by _Eng and _Tel correspond to English and Telugu words respectively. Interestingly, as observed in the example, code-mixing occurs at morpheme level in case of 'subjectlu': subject (English root) + 'lu' (Telugu plural morph inflection). In the above example, the dominant language is Telugu, which is the language into which certain words are mixed, also known as the *matrix language*. The other language, whose constituents are brought into the *matrix language* is called the *embedded language*, which is English in this case.

Understanding Dialog Acts (DA)s is a very essential topic in progressing towards conversational analysis. The origin of dialog acts traces back to Austins theory [2] of locutionary, illocutionary, and perlocutionary acts. Searle has introduced the concept of speech acts (assertives, directives, commissives, expressives, declarations) [3], that come under Austins illocutionary acts. DAs are special kinds of speech acts and their granularity is largely dependent on the domain.

Similar to general scenarios, the structure in code-mixed conversations can be studied by representing it as consecutive adjacency pairs like (Question, Answer), (Offer, Accept), (Greeting, Greeting), (Compliment, Acknowledge) etc,. The context in which the sentence is embedded plays an important role in understanding the intent of the sentence. For example, consider these sentences that are used in day-to-day common language: 'salt ni pass cheyagalava please' *(Translation: Please pass the salt)* and 'neeku Hindi lo poem cheppadam vacha?' *(Translation: Can you tell a poem in Hindi?)*. The current approaches of dialog act tagging are not equipped enough to handle such switch in languages, especially when the switching juncture point is unclear. Hence we are using existing low level resources to address this problem to correctly understand that the former is an action directive and the latter is an information request.

In this paper, we take a lexical translation based approach using an at home developed language identification model and we put different learning algorithms into use and analyze the results.To the best of our knowledge, we are the first to introduce this problem and to come up with

[1] http://chaibisket.com/

Speaker	Utterance	Translation
SYS:	Hi nenu mee Library Assistant	I am your Library Assistant
	meeku ela help cheya galanu ?	How can I help you ?
USER:	Hello	Hello
	linear algebra books section ekada undi ?	Where is linear algebra books section ?
SYS:	Linear Algebra books meeku section 3.2 lo dorukutundi	You will find Linear Algebra books in section 3.2
	meeku ye book kavali	Which book do you want ?
USER:	Naaku Linear Algebra by Russel norvig third edition kavali.	I want the third edition of Linear Algebra by Russel norvig

Table I: Sample Data

strategies to deal with this real world language dynamics.

This paper is divided into five sections. We first discuss the related work in this field. Section 3 discusses our data collection method and data statistics. Section 4 describes our approaches for DA tagging along with experimental results. Section 5 concludes and discusses issues.

table

II. Related Work

There are a few latest advancements that happened in code-mixing like development of POS taggers [4], approaches towards building a shallow parsing pipeline [5], building a Question classification system for Code-Mixed Questions [6]. But to the best of our knowledge, there is no work that discusses code-mixing in a dialog scenario, where recognizing the intention of a speaker is of primary importance.

DAs can be used for the purpose of intention recognition in a task oriented dialog system. There are a number of significant contributions in the area of Dialog Act recognition [7]. Arabic and Spanish are some other languages in which some work on DA annotation and recognition has been published [8], [9]. Also semi supervised Dialog Act tagging approach has been proposed for Telugu [10].The data set used has very short dialogs, atmost two dialog acts per conversation. This short a dialog cannot be used for effective conversational analysis. Many annotation schemes have been developed from projects like DAMSL, a domain independent DA annotation schema, [11], Maptask [12] and Verbmobil [13]. The previous work in this domain is performed on corpora that is not fully representative enough of the casual day-to-day conversations in a multilingual environment.

III. Data Collection : Wizard of Oz

To the best of our knowledge there is no code-mixed conversational data publicly available for any language pair. Therefore, English-Telugu conversational data was collected through WOz experiment. 28 participants are involved in this including students and faculty (library staff) of an educational institute. Out of these 3 people assumed the role of a virtual library assistant (a computer system) and the remaining 25 interacted with the wizards with various queries, resulting in 25 conversations. To ensure diversity within the domain, participants were not provided with specific tasks and were free to ask any questions pertaining to library.

Initially, they were asked to use Telugu, but most of the participants started mixing English. When asked to stick to a single language (Telugu), the response time of the participant increased, and the conversation lost its naturalness. So, this strict imposition was removed and instead we asked them to speak to the system in a natural way as how they would pose the question to another peer who knows Telugu. Along with *code-mixing*[2], some participants also did *code-switching* [3]. So, in a given turn of a speaker, there can be three possibilities: An utterance is either completely in English, Telugu, or is code-mixed. Resembling cross-scripting observed in social media like Facebook, Twitter and massive usage of English keyboards in Romanizing native languages, this data is also collected in similar cross-scripted manner via a chat interface. Table 1 depicts an excerpt from a conversation collected through WOz. A total of 25 conversations were collected and Table 2 shows statistics of the data.

Total no. of conversations	25
Total no. of participants per conversation	2
Total no. of utterances	856
Total no. of unique utterances	636
Total no. of words	4,270
Total no. of unique words	1,147
Average no. of turns/conversation	21
Average no. of utterances/conversation	34
Average no. of words/conversation	171
Average no. of words/utterance	5
Average no. of utterances/turn	2-3
English Utterances	172
Telugu Utterances	93
Code-mixed Utterances	352

Table II: Data Statistics

Dialog Act Annotation: The data was annotated by two annotators in two dimensions, given the DAMSL guidelines: forward communicative function and backward communicative function. A 'NULL' tag is included in both Backward and Forward Looking Functions, which marks the absence of a relation between the current speaker's utterance and the previous speaker's utterance and in cases of a grounding utterance. For example an utterance like 'okay' is tagged as 'NULL' in the Forward Looking layer. Cohen's kappa was calculated and an inter-annotator agreement of 87.19 has been obtained, which is a sign

[2]Code-Mixing is the embedding of linguistic units such as phrases, words and morphemes of one language into an utterance of another language.

[3]Code-Switching is juxtaposition within the same speech exchange of passages of speech belonging to two different grammatical systems or sub-systems

of reliable data. Figure 1 shows the counts of different forward and backward tags that are annotated in the data.

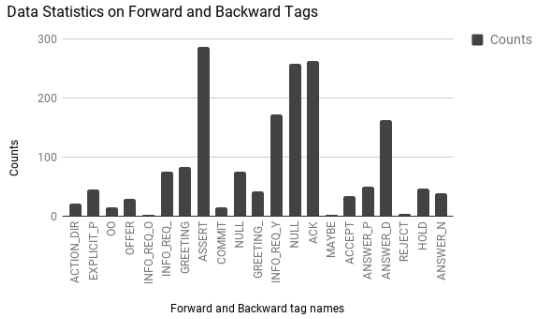

Figure 1: Data statistics on Forward and backward tags

table

IV. Our Approach

In this section, we give an outline of the pipeline of our system.

A. Data Pre-Processing

A speaker could utter more than one sentence per turn and some utterances could be shorter than a sentence (for example: 'Sare'(Okay) and 'thappakunda'(Sure)). In chat text, multiple punctuation marks could be an indication to a pause.For the task of tokenization, NLTK sentence and word tokenizers have been deployed. In the following example, punctuation is used to signify pause, but sentence tokenizer would separate them into two different sentences.

Before Segmentation: *USER: ikkada issue cheste...mari nenu eppudu collect chesukovali?*

After Segmentation: *USER: ikkada issue cheste mari nenu eppudu collect chesukovali?*

Hence we have manually checked for such instances.

B. Language Identification(LID)

We have used the data from *'Chaibisket'* website to develop a language identification model. The crawled data has been annotated with six categories, namely *English, Telugu, Mixed (morpheme level mixing), acronyms, named-entities and unknown words.* We have experimented with standard learning algorithms like SVM, KNN, CRF and Neural Networks. The surface level lexical and sub-lexical feature set used for this task comprises of lexicon, prefix, suffix, infix, presence of post positions, length of the word, neighbouring words, emoticons, alphanumeric characters, casing. Comparatively, CRF performed the best by using this feature set with a precision of 90% ,recall of 91% and an F-score of 88%. The accuracy of the system is 90.6%.

C. Transliteration and translation

In addition to the dearth of code-mixed data, the problem of learning reliable word embeddings is further compounded with spelling variations due to Romanization of Telugu words. In our approach, we investigate if translation could be of any use. The output from LID system is used for this purpose. At this stage, we have two options, either translate everything in to English or everything into Telugu. We investigate either ways. We have used Indictrans for transliteration and Google translate for lexical(individual word) translation.

D. Learning Algorithms

We have explored multiple learning algorithms namely, SVM, KNN, HMM, Naive Bayes, MLP and LSTM and have tried to understand how differently code-mixed data needs to be processed as compared to monolingual data.We began with exploring traditional machine learning algorithms for code-mixed data. We model the DA Tagging problem as a sequence labelling problem and therefore, used an HMM and also an LSTM in the later experiments. To combat the problem of lack of annotated data for the task at hand, we are using sliding window based splits in the conversations for sequence labeling approaches like HMM and LSTM.

Besides, individual unigrams, bigrams and trigrams,a weighted combination of grams- assigning higher weight to trigrams compared to bigrams and unigrams are used for featurizing SVM, KNN, Naive-Bayes and HMM. Whereas, the features used for MLP and LSTM are word-embeddings.

In case of no translation, words in a single sentence occur from both English and Telugu which does not have a common vector space (explained in section 4.4). Therefore word-embeddings cannot be used without translating into either of matrix or embedded languages. Hence MLP and LSTM are not used on pre-translation CM data.

Table 3 comprehensively presents the results obtained through various experiments.

V. Discussion and Conclusion

As observed from Table 3, HMM performs better comparatively, the reason for this being the inclusion of contextual information. The translation has not added significant value to the scores, the reasons for which are discussed here. When Telugu words in a code-mixed sentence are transliterated and then translated to English, errors can occur at both these levels, where complete loss of original word and insertion of wrong word occur respectively. This might result in not finding corresponding word vectors.

There are no strict guidelines that need to be adhered while Romanizing Telugu. This leads to spelling variations. For example, consider the words: 'vachaadu' and 'vachadu' (wx form: vaCADu, meaning: 'he came'). Another most commonly found such variations is the presence and absence of 'h'. This heavily depends on idiolect as compared to merely geographical or societal factors. Telugu is an agglutinative language which implies that it is rich in its morphological structures. As a result, the transliteration and translation may not be accurately available for combinatorial words. For example, 'neek-enduku' (neeku + enduku). The issue here is that both of the individual words are present in dictionary but not the combined word. The issues about the shortened writings of words in informal settings, where standardization is

Learning Algorithm	Translation	Function	Best feature	Precision	Recall	F-score
SVM	NO	Forward	unigrams	12.94	25.54	15.44
		Backward	unigrams	21.95	32.58	24.23
KNN	NO	Forward	unigrams	50.76	46.35	46.15
		Backward	unigrams	49.32	47.42	45.16
HMM	NO	Forward	unigrams	88.43	82.16	**82.5**
		Backward	unigrams	79.89	71.44	**70.77**
Naive Bayes	NO	Forward	unigrams	85.35	75.53	77.95
		Backward	unigrams	67.92	57.42	59.22
HMM	YES	Forward	unigrams	88.27	82.00	**82.34**
		Backward	unigrams	80.61	71.56	**70.99**
Naive Bayes	YES	Forward	unigrams	85.28	74.67	76.88
		Backward	unigrams	69.28	56.41	58.66
MLP	YES	Forward	word_embeddings	71	69	69
		Backward	word_embeddings	67	65	65
LSTM	YES	Forward	word_embeddings	28	38	32
		Backward	word_embeddings	38	39	38

Table III: Results obtained for DA tagging for CM utterances with different algorithms

All tokens	26034
English tokens	11275
Telugu tokens	10599
Universals(numbers,punctuation)	3435
Named entities	575
Acronyms	69
Other	78

Table IV: Statistics of Chai-Bisket Data

demanded before further processing. For example, in chat and SMS language, often 'u' maps to 'you', 'cooool' maps to 'cool' respectively, with variable number of 'o's.

Translation into Telugu was comparatively poor.Romanized Telugu words do not adhere to specified rules, hence introducing errors in translation. While translating to English, words identified as English remain the same, thus eliminating translation error.

Lexical translation has been effective in dealing with CM as shown by [6] where translation improved accuracy of their question classification system by 5%. We have not used topic-specific features to maintain the generalizability of the system across domains. [14] claims that maximum-entropy approach combined with Neural-Networks(MLP) gave the lowest error rate with an accuracy of 78.7 on ICSI meeting corpus. We speculate that this implementation for our CM setting would improve the F-score. However, the unnormalized spellings of Romanized Telugu words pose challenges in creating efficient word-embeddings.

Our best performing system is HMM based which has given an average F-score of 76.67, averaged over Forward and Backward functions. Our future work is in the direction of addressing these issues to build reliable DA tagger that can be used in day-to-day chat conversations.

REFERENCES

[1] T. Hidayat, "An analysis of code switching used by facebookers," 2008.

[2] J. L. Austin, *How to do things with words*. Oxford university press, 1975.

[3] J. R. Searle, *Speech acts: An essay in the philosophy of language*. Cambridge university press, 1969, vol. 626.

[4] A. Jamatia, B. Gambäck, and A. Das, "Part-of-speech tagging for code-mixed english-hindi twitter and facebook chat messages." Association for Computational Linguistics, 2015.

[5] A. Sharma, S. Gupta, R. Motlani, P. Bansal, M. Srivastava, R. Mamidi, and D. M. Sharma, "Shallow parsing pipeline for hindi-english code-mixed social media text," *arXiv preprint arXiv:1604.03136*, 2016.

[6] K. C. Raghavi, M. K. Chinnakotla, and M. Shrivastava, "Answer ka type kya he?: Learning to classify questions in code-mixed language," in *Proceedings of the 24th International Conference on World Wide Web*. ACM, 2015, pp. 853–858.

[7] A. Stolcke, K. Ries, N. Coccaro, E. Shriberg, R. Bates, D. Jurafsky, P. Taylor, R. Martin, C. Van Ess-Dykema, and M. Meteer, "Dialogue act modeling for automatic tagging and recognition of conversational speech," *Computational linguistics*, vol. 26, no. 3, pp. 339–373, 2000.

[8] S. B. Dbabis, H. Ghorbel, L. H. Belguith, and M. Kallel, "Automatic dialogue act annotation within arabic debates," in *International Conference on Intelligent Text Processing and Computational Linguistics*. Springer, 2015, pp. 467–478.

[9] K. Ries, "Hmm and neural network based speech act detection," in *Acoustics, Speech, and Signal Processing, 1999. Proceedings., 1999 IEEE International Conference on*, vol. 1. IEEE, 1999, pp. 497–500.

[10] S. Dowlagar and R. Mamidi, "A semi supervised dialog act tagging for telugu," *ICON*, 2015.

[11] J. Allen and M. Core, "Draft of damsl: Dialog act markup in several layers," 1997.

[12] A. H. Anderson, M. Bader, E. G. Bard, E. Boyle, G. Doherty, S. Garrod, S. Isard, J. Kowtko, J. McAllister, J. Miller *et al.*, "The hcrc map task corpus," *Language and speech*, vol. 34, no. 4, pp. 351–366, 1991.

[13] J. Carletta, S. Isard, G. Doherty-Sneddon, A. Isard, J. C. Kowtko, and A. H. Anderson, "The reliability of a dialogue structure coding scheme," *Computational linguistics*, vol. 23, no. 1, pp. 13–31, 1997.

[14] M. Zimmermann, D. Hakkani-Tür, E. Shriberg, and A. Stolcke, "Text based dialog act classification for multiparty meetings," in *International Workshop on Machine Learning for Multimodal Interaction*. Springer, 2006, pp. 190–199.