Discovering Language Independent Latent Aspect Clusters from Code-Mixed Social Media Text

Kavita Asnani
Goa College of Engineering
Goa, India.
kavita@gec.ac.in

Jyoti D. Pawar
Goa University
Goa, India.
jyotidpawar@gmail.com

Abstract

In recent times, code-mixing has become prevalent in social networking as people communicate in multiple languages. This is become a trend and is significantly popular especially in multilingual countries. This has led to the generation of large code-mixed text having useful topics of information dispersed. However, it is very challenging as the code-mixed social media text suffers from its associated linguistic complexities. The main focus of this work is discovery of latent topics indicating useful information from code-mixed social media text overcoming the barriers of random language switch. We evaluate the resulting topic aspect clusters on standard lexical semantic evaluation tasks and show that our method produces substantially better semantic representations than code-mixed counter parts.

Introduction

Communication over social networking forums has become indispensable for users as they actively use it for routine communication. With the growing popularity of social media, recent research has concentrated on English chat data or on multilingual data where mixing occurs at sentence level. However, increasing large volumes of code-mixed messages contain useful information dispersed in high dimensional unstructured text. The task of discovering the significant information from such data is challenging due to the following inherent characteristics of code-mixed data:

- Short and long length messages: Social media code-mixed text comprises of informal communication with random mix of words in different languages in short and long messages, thereby distributing semantics across varied vocabulary.
- Noisy: The chat style of communication involves use of slang and ungrammatical words (Dhuliawala, Kanojia, and Bhattacharyya 2016).
- Codemixing at different levels of code-complexity:Code-Mixing in social media text occurs with random mix of words in different languages measured by the codemixing index measure (Das and Gambäck 2014) based on the number of languages participating in mixing.
- Language Detection: Since codemixed data consists of mix of words occurring in different languages, semantic interpretation of such data is possible only by identifying the language in which words are written (Solorio et al. 2014). Several explicit language identification systems need to be used for this purpose.

Therefore, code-mixed content exhibits most of the language related problems. At this point, it is worth noting two points. First, though a code-mixed message contains random mix of multiple languages in the same sentence, it still has a valid meaning, as words in different languages refer to the same context. Second, code-mixed social media data refer to informal chat communication among people and thus it contains diversity of topics that people care about. We follow the claim in (Heinrich 2009) which stated that the words occurring in the same context tend to be semantically similar. Hence, the main motivation for carrying out this research is discovery of useful and coherent information from code-mixed social media data spread across random mix of languages, by means of aspect semantic clusters obtained using topic modeling. Code-mixed social media content contains ungrammatical words which needed pre-processing to be done. We found Shallow parser (Sharma et al. 2016) to be suitable for this purpose to obtain normalized output. In this paper, we address our objective using a two-step procedure on pre-processed code-mixed data for discovering aspect topic clusters. The first step focuses exclusively on considering each code-mixed message as a document by updating the core representation of the code-mixed document in monolingual form. This paper points to the fact that semantic interpretation of code-mix content is invariant across random mix of languages and proposes a simple technique which works by leveraging knowledge from multilingual lexicographic semantic resource called BabelNet(Navigli and Ponzetto 2012) for incorporating monolingual representation. Second, in order to retrieve useful latent aspects of topic clusters, our proposed approach attempts to use unsupervised probabilistic topic models.

The remainder of this paper is organized as follows: Section 2 presents related work; Section 3 describes our proposed model; Section 4 gives implementation details; Section 5 presents experimental results and Section 6 states conclusion.

Related Work

Recently code-mixing has been popularly observed on social networking forums. India is a multilingual country. Across 29 states with over 22 official languages, millions of people communicate over social networks for routine tasks. Such high level of language diversity encourages frequent code-mixing in social media context. In general, multilingual social media users communicate in two languages. In (Sharma, Choudhury, and Vyas 2014), the analysis of code-mixed bilingual English-Hindi data showed significant amount of code-mixing and proved that most of the active users on Facebook are bilingual.However, context in the code-mixed data is shared across random occurrence of words across different languages.(Harris 1954) stated distributional hypothesis structure which says that the meaning of a word is based on the context shared by the words which co-occur. Therefore, we performed the survey of work related to language identification at the word-level. (Solorio et al. 2014) contributes to the first shared task on language identification in code-switched social media twitter data which included data from four language pairs namely Modern Standard Arabic-Dialectal Arabic (MSA-DA); Mandarin-English (MAN-EN), Nepali-English (NEP-EN), and Spanish-English (SPA-EN). The evaluation results with the lowest F-measure of the task showed that language identification at token level is more difficult when the languages present are closely related as in MSA-DA. In our case, the semantics present in the underlying context of code-mixed communication is expressed across random occurrences of code-mixed words. Also, the use of machine translators is very challenging and not feasible as the volume of the data is large, inconsistent and translation is required to be done at the word-level. Common methods using parallel or comparable corpora for machine translation, does not work for code-mixed social media data due to the gap in terminology, domain mismatch and unavailability of training data for various combination of languages. To address these shortcomings, we propose language independent representation of code-mixed messages. Using a shallow parser (Sharma et al. 2016), our proposed approach first addresses noise elimination and need for normalization. We augment each code-mixed word with its semantic complement represented in corresponding languages used for random mixing in the code-mix content, thereby dropping the language barrier. For this purpose, we leverage knowledge from BabelNet (Navigli and Ponzetto 2012) as it provides the same concept expressed in many different languages. BabelNet 3.6 is a wide-coverage multilingual semantic network created from integration of both lexicographic and encyclopedic coverage of terms. The integration is performed by an automatic linking method and lexical gaps are filled with the aid of machine translation. As a result, multilingual concepts and named entities are connected with numerous semantic relations and are provided with as many as 14 million entries called Babel synsets. Such multilingual synsets offer synonyms in range of different languages representing a given meaning. In our work we chose HTTP RESTful Java API of BabelNet v3.6 which covers 271 languages and is obtained from automatic integration of WordNet, Open Multilingual WordNet, Wikipedia, Omega Wiki, Wikitionary and Wikidata. As we are interested in using large collection of code-mixed chat documents and deriving useful aspects of text in the form of clusters of similar themes together regarded as topics; so we use unsupervised topic model. Such clusters offer useful information of significant topic aspects from social media communication to the administrator or end user. To meet our objective, we used Probabilistic Latent Semantic Analysis (pLSA) (Hofmann 1999) and Latent Dirichlet Allocation (LDA) (Blei, Ng, and Jordan 2003) as they are highly recommended unsupervised topic modeling methods for this purpose.

Our Proposed Work

We first present how we addressed random occurrence of words in a code-mixed message. In order to automatically deal with words occurring in different languages in the code-mixed text input, we construct augmented sets by directly obtaining semantic interpretations of words in the form of synsets from BabelNet in the corresponding language used in mixing. But for each code-mixed word it resulted in retrieval of large number of synsets. This is due to all the possible multilingual synonyms assigned at the word level. Therefore, for correct semantic interpretation we had to ensure that the synsets are mapped based on the part of speech (POS) tags and the complement language used for code-mixing. Such synsets in the form of augmented sets addressed the monolingual vocabulary across languages.

The Code-mixed Word Augmentation Process

Since our aim is language independent representation of codemixed chat messages, we treat each message independently and divide it into stream of words.Given collection of codemixed messages in |L|languages where L=$\{l_1, l_2, l_{3,...}, l_l\}$ is set of languages in which code-mixing has occurred. The codemixed message collection is denoted as M =$m_1^L, m_2^L, m_3^L ..., m_n^L$ where n denotes number of codemixed messages in L languages. Each code-mixed message is represented as $m_i^L = x_1^{l_i}, x_2^{l_i}, x_3^{l_i} ..., x_{|m|}^{l_i}$ where $x_i \in V_L$ $w_{i1}^{l_i}, w_{i2}^{l_i}, w_{i3}^{l_i}, ..., w_{iN_i}^{l_i}$ is the word in the vocabulary across the chat. Here, x_{i}^l denotes the i^{th} word in a code-mixed corpus in a certain language l_{i} where $l_{\text{i}} \in$ L and N_i denotes number of words in the i^{th} message and w_{ij} denotes j^{th} word of the i^{th} message. Our augmented code-mixed data therefore contains the revised and extended vocabulary. We found this knowledge to be beneficial to our further employed topic models as each topic is a multinomial distribution over augmented sets. We propose to provide our codemixed and monolingual corpora further to two topic models. First, we chose to use codemixed PLSA proposed by (Asnani and Pawar 2016) which is based on co-occurrence matrix for codemixed message which is modeled on PLSA to discover topic aspects. We refer to CodeMixed PLSA as CM-PLSA. Second, we propose to use LDA (Blei, Ng, and Jordan 2003) for topic discovery.The codemixed and language independent code-mixed topics across the chat collection across respective topic clusters is given as: Z= {

```
foreach w_i ∈ m do
    | w_l ← list {lang(w_i), pos(w_i )}
    end
for i = 1 to n do
    foreach t in w_l do
        if lang(t) = L then
            | S_L ← synset_L(t)
        else if lang(t)= L' then
            | S_L ← translatedSynset(t, L)
        foreach synset s in S_L do
            if pos(s) = pos(t) then
                | t ← t + s
        end
    end
end
```

Algorithm 1: Code-mixed Word Augmentation Process

$z_1, z_2, z_{3,...}, z_k$ }. Fig. 1 shows the sample clusters generated by CM-PLSA for codemixed and monolingual corpora.We have presented the augmentation process in Algorithm 1.

Input : Tagged code-mixed message,
Codemix Language1: L1,
Codemix Language2: L2
Output : Annnotated codemix message

Implementation Details

In our experiments, our proposed approach can deal with words randomly mixed in either of the two languages Hindi or English. We compared the models by measuring coherence of aspect topics. For evaluating quality of topics we used KL-Divergence and topic coherence(UMass). We tested with different number of topics z and compared topic aspect clusters of the same size.

Dataset Used

We performed experiments on FIRE 2014[1](Forum for IR Evaluation) for shared task on transliterated search. This dataset comprises of social media posts in English mixed with six other Indian languages.The English-Hindi corpora from FIRE 2014 was introduced by (Das and Gambäck 2014). It consists of 700 messages with the total of 23,967 words which were taken from Facebook chat group for Indian University students and it contained 63.33% of tokens in Hindi. The code-mixing percentage for English-Hindi corpus was as high as 80% due to the slang used in two languages randomly during the chat. We then used POS tagger by (Petrov, Das, and McDonald 2011) to obtain POS tag of each word.

[1]http://www.isical.ac.in/ fire/

Experimental Results

Measuring Topic Quality by Topic Distinctiveness

The Kullback Leibler (KL) divergence measure (Johnson and Sinanovic 2001) is a standard measure for comparing distributions. We applied symmetrical version of KL Divergence and averaged it across the topics. We analyzed distinctiveness of the topic aspects produced by CM-PLSA and LDA methods as we were interested in measuring the overlap across the topics based on the co-occurrence of the terms in multiple topic clusters and tendency for a method to generate distinct topics from the respective code-mixed and monolingual data. The average KL Divergence score thereby for all the clusters was calculated and the results for values $z \in [3, 6, 9, 12]$ are presented in Fig. 2 and Fig. 3 respectively. The overlap in topic aspects produced by both the methods is high for topic size $z = 3$ across the code-mixed and monolingual corpora due to lower number of unique terms across the topics. In Fig. 2 and Fig. 3 we can see pattern for $z \in [6, 9, 12]$, where it can be seen that monolingual Hindi (Mono-Hin) corpora generates high KL-Divergence score across both the methods due to the high probability of unique terms and also we attribute this observation to the fact that basic 63.33% of Hindi tokens participated in the code complexity of the core code-mixed input. We observe consistent drop in distinctiveness due to increasing number of general terms that may be discriminating.

Measuring Topic Quality by Topic Coherence

We analyzed topic quality for measuring its interpretability using UMass score, which is a pairwise score function of the empirical probability of common words (Mimno et al. 2011). Average UMass coherence score for code-mixed and monolingual corpora for both the CM-PLSA and LDA in Fig. 4 and Fig. 5 respectively. Our results show that Augm-CM have high UMass scores for topic size z=3, as more of words describing the term relating to the same underlying semantic are explicitly specified in the data input. The pattern drops with increasing topic size. However, topic interpretability is more stable for different number of topics $z \in [3, 6, 9, 12]$. It is relatively on the higher side for Mono-Eng and Mono-Hin data forms of the corpora showing better topic quality.

Conclusion

We proposed novel approach which utilizes semantic knowledge freely available in external multilingual resource which can drop the language barrier from code-mixed data resulting in discovery of useful topic aspects. Our evaluation results conclude that our approach is effective in generating highly distinctive aspect topic clusters where aspect clusters inferred from monolingual representation of data are consistently coherent as compared to those obtained from their core code-mixed counterparts.

References

Asnani, K., and Pawar, J. D. 2016. Discovering thematic knowledge from code-mixed chat messages using topic model. *Proceeding of the 3rd WILDRE, organized with LREC, Portorož, Slovenia.*.

CM	Augm-CM	Mono-Eng	Mono-Hin
didnt 7.529599735e-03	बयान 6.847152571e-03	side 9.895503103e-03	जानना 1.127027921e-02
उस्क 7.508342752e-03	तुने 6.845717272e-03	make 9.895503103e-03	परमाणु_संख्या_7 1.1277921e-02
good 7.508342752e-03	करते 6.831249903e-03	male 9.895503103e-03	अभी_भी 1.123888234e-02
english 5.136567256e-03	तेरी 5.197220689e-03	future 9.854387535e-03	तोह 7.740204781e-03
students 5.112360007e-03	होती 5.181658418e-03	fmly 9.854387535e-03	स्थिति 7.652304615e-03
करो 5.109157680e-03	मैच 5.165837751e-03	love 5.326374453e-03	वह 7.651188060e-03
talk 5.088062458e-03	तोह 3.537829441e-03	life 5.216837385e-03	एएन 7.646460742e-03
ignoring 5.088062458e-03	बर 3.516676055e-03	time 5.206495402e-03	पक्ष 7.615301250e-03
waiting 5.088062458e-03	क्रने 3.516676055e-03	home 5.206495402e-03	समय 1 7.615301250e-03
senior 5.088062458e-03	लग्ता 3.514509176e-03	face 5.206495402e-03	कॉन्सर्ट 7.615301250e-03
दूसरी 5.088062458e-03	dad 3.514509176e-03	hav 5.206495402e-03	जीत 7.615301250e-03

Figure 1: Sample Aspect Topics with Probability (Top n probability aspects comprise topics)

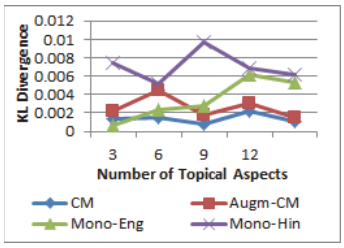

Figure 2: KL:CM-PLSA

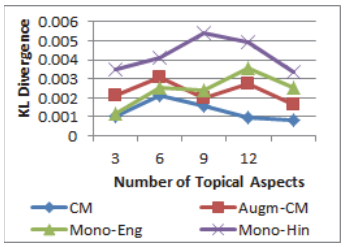

Figure 3: KL: LDA

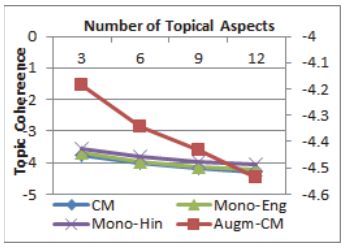

Figure 4: UMass:CMPLSA

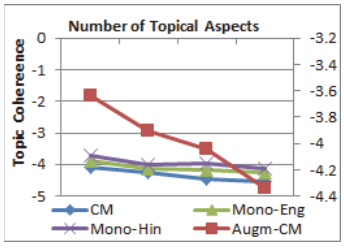

Figure 5: UMass: LDA

Blei, D. M.; Ng, A. Y.; and Jordan, M. I. 2003. Latent dirichlet allocation. *Journal of machine Learning research* 3(Jan):993–1022.

Das, A., and Gambäck, B. 2014. Identifying languages at the word level in code-mixed indian social media text. *11th International Conference on Natural Language Processing (ICON-2014)*.

Dhuliawala, S.; Kanojia, D.; and Bhattacharyya, P. 2016. Slangnet: A wordnet like resource for english slang. In *Language Resources and Evaluation Conference (LREC 2016)*, volume 10. ACL.

Gambäck, B., and Das, A. 2016. Comparing the level of code-switching in corpora. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, 1850–1855.

Harris, Z. S. 1954. Distributional structure. *Word* 10(2-3):146–162.

Heinrich, G. 2009. A generic approach to topic models. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, 517–532. Springer.

Hofmann, T. 1999. Probabilistic latent semantic analysis. In *Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence*, 289–296. Morgan Kaufmann Publishers Inc.

Johnson, D., and Sinanovic, S. 2001. Symmetrizing the kullback-leibler distance. *IEEE Transactions on Information Theory*.

Mimno, D.; Wallach, H. M.; Talley, E.; Leenders, M.; and McCallum, A. 2011. Optimizing semantic coherence in topic models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 262–272. Association for Computational Linguistics.

Navigli, R., and Ponzetto, S. P. 2012. Babelnet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence* 193:217–250.

Petrov, S.; Das, D.; and McDonald, R. 2011. A universal part-of-speech tagset. *arXiv preprint arXiv:1104.2086*.

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

Sharma, K. B. J.; Choudhury, M.; and Vyas, Y. 2014. i am borrowing ya mixing? an analysis of english-hindi code mixing in facebook. *EMNLP 2014* 116.

Solorio, T.; Blair, E.; Maharjan, S.; Bethard, S.; Diab, M.; Gohneim, M.; Hawwari, A.; AlGhamdi, F.; Hirschberg, J.; Chang, A.; et al. 2014. Overview for the first shared task on language identification in code-switched data. In *Proceedings of The First Workshop on Computational Approaches to Code Switching*, 62–72. Citeseer.