A Survey of Current Datasets for Code-Switching Research

Navya Jose
Machine Intelligence
Indian Institute of Information Technology and Management-Kerala
Trivandrum, India
navya.mi3@iiitmk.ac.in

Bharathi Raja Chakravarthi, Shardul Suryawanshi *
Data Science Institute
National University of Ireland
Galway, Ireland
bharathi.raja, shardul.suryawanshi@insight-centre.org

Elizabeth Sherly
Machine Intelligence
Indian Institute of Information Technology and Management-Kerala
Trivandrum, India
sherly@iiitmk.ac.in

John P. McCrae*
Data Science Institute
National University of Ireland
Galway, Ireland
John.McCrae@insight-centre.org

***Abstract*—Code switching is a prevalent phenomenon in the multilingual community and social media interaction. In the past ten years, we have witnessed an explosion of code switched data in the social media that brings together languages from low resourced languages to high resourced languages in the same text, sometimes written in a non-native script. This increases the demand for processing code-switched data to assist users in various natural language processing tasks such as part-of-speech tagging, named entity recognition, sentiment analysis, conversational systems, and machine translation, etc. The available corpora for code switching research played a major role in advancing this area of research. In this paper, we propose a set of quality metrics to evaluate the dataset and categorize them accordingly.**

***Index Terms*—code switching, natural language processing, dataset**

I. Introduction

With the advent of social media, users prefer to mix multiple languages in their online platform conversations [1]. The reality is that the upcoming generations will fluently speak more than one language and hence we could say the future belongs to multilingual speakers. This has been accompanied by a myriad of trends among the users. It has been observed that people fluent with multiple languages switch between them to compensate for the deficiency of expressions [2], [3]. It is found that the extensive use of code-switching enabled multilingual speakers to express their idea more precisely and convincingly [4]. When the speaker knows multiple languages, it is a quite common act to switching from one language to the other. This back and forth switching of languages in the same conversation between multilingual speakers is known as code-switching [5], [6]. This is a natural conflation between people with more than one language in common and this trend has been extensively found in the multilingual communities. Many languages in the world have their scripts [7], [8] but code-switching in the online text forums Roman script is used most frequently for convenience of typing which paved the way for new era code-switching in online textual communication [9]–[11]. Due to this kind of change of words or sentences between different languages in a single conversation, it becomes difficult to find out the languages involved [12].

In this new era of various modes of communication, a major portion of the data accessible today is unstructured and most of it also contains code-switched data. The performance of the traditional natural language processing techniques of this mixed data is unsatisfactory even at a basic natural language processing such as language identification at the word level. Creating new methods and tools for natural language processing tasks of code-switched data suffer from the unavailability of data in this domain. In this paper, we propose a set of quality metrics to evaluate the code-mixed corpus and list out the available corpora for different language sets.

II. Background

The term code-switching has not coined in the early years of the twentieth century, however, code-switching in textual form become more prominent after the Internet boom. This was considered only as of the imperfection in the language learning process [13], it left unattended by the researchers for a long time. Recently many studies have been done to crack code-switching data to automatically process such content from social media. One of the usages would be to detect the offensive content in the media and remove it. Even if the natives use their script instead of Roman script, the dialects cause issues regarding identification, translation and further processes. As there are so many languages in the world with or without a script. Illustrious examples available for our topic of interest is from public networking sites. Here specifically the code-switched texts of Indian languages and English are taken for illustration.

*Authors are supported by a research grant from Science Foundation Ireland, co-funded by the European Regional Development Fund, for the Insight Centre under Grant Number SFI/12/RC/2289.

978-1-7281-5197-7/20/$31.00 ©2020 IEEE

A. Code-Switching

Code-Switching is switching between languages within a single context, it may occur in a sentence by sentence basis or in the same sentence. It is not merely a matter of knowledge in more than one language. There is no point if the speaker is not able to follow the constraints of the grammar of those languages. While using any language, there is a trend of borrowing from other languages. If a word once used in between another language and if repeated henceforth, that word is called borrowed word [14]. Borrowed words are therefore added to the lexicon of a new language. Code-switching is different from borrowing since it keeps the identity of each language though used frequently in between. Social media text is complicated due to the wider use of acronyms and code-switching makes it even more difficult to process.

The custom of altering between two or more languages in the same sentence is referred to as code-switching. Code-switching enables the user to continue the conversation without any sort of interference or barriers of a single language. It is true that when upset or distracted, one can express their exact feelings in the native language better foreign language. There is a choice of words to express the emotions in such instances as well as this enables the speaker to cause or impact on the listener. Hence it is advantageous for communication [15].

The first step of any linguistic study is language identification.

Njan ente frndne help cheyyukayayirunnu.

I was helping my friend.

Shown above in italics is an example of code-switching and the corresponding English gloss is also depicted. The words 'frnd'(friend) and help is from English words mixed with Malayalam (a south Indian language).

B. Types of Code-switching

Earlier by Gumperz(1972) [16] it was claimed that there are mainly two types of code-switching such as situational and metaphorical. Code-switching is *situational* when the speaker switches the topic into another situation in different languages in between the same conversation. When this change of language is to add another mood to the words distinct from the rest of the conversation like stressing or softening is said to be *Metaphorical*. The following is the example of situational and metaphorical code-switching. Corresponding English elucidation is given below each.

1) *aaj ki pariksha kal ki pariksha se behtar dhi.Btw wen will u come home?*
   Today's exam was better than yesterday's. by the way when will you come home?
2) *Aaj tumhari class hogi.All should be there on time.*
   you will be having class today. all should be there on time.

The sample text shown above is to illustrate the difference between situational and metaphorical code-switching. Both are a mixture of English and Hindi (Indian language). In the first example, switching depicts the change of situation. In the second example, it is the teacher's message to the students and the importance of the speaker and message is well expressed by switching to English (the more formal language). Another categorization is into inter-sentential and intra-sentential [17]. Despite the fact that code mixing is considered to be the synonym of code switching, the former refers to the alteration of two or more languages in the same sentence *(intra-sentential code-switching)* and later allude to the same act between sentences*(inter-sentential code-switching)* [18]. For example:

1) *nee naale evng shoppingnu varunnundo?.*
   Are you coming for shopping tomorrow evening?
2) *Exactly. Unakkepadi theriyum?*
   Exactly. How do you know this?

The sentences quoted in italics are a mix of English with Malayalam and Tamil respectively, the languages spoken in the southern region of India. The first one is an example for intra-sentential-code-switching as the mixing of English words is done inside the sentence. The other one shows inter-sentential-code-switching where the speaker alternates language between sentences.

III. Quality Criteria for Code Switching Dataset

The quality of dataset highly depend on the data collection techniques and annotation methods. Most of the dataset were scrapped or crawled from social media and annotated by expert or crowd-sourcing. The quality of data in the early stages plays a major role in the research. However, content of the data can play role in narrowing down the research. Code-Switching got attention recently, the data may not be available for all the research area of Natural Language Processing. In this section, we propose a set of metrics to characterize the code-switched dataset.

A. Number of Words

The number of words in the corpus shows us the size of the corpus. For language identification, named entity recognition, and POS tagging problem the number of words would give more intuition about the corpus.

B. Vocabulary Size

One of the main difficulties in training the advanced deep learning model is the computational complexity of computing the target word in language model and machine translation, this is Out-of-Vocabulary (OOV) problem. That is unseen words during training time. Since code-switching in social media contain non-standards spelling of the word and written in different scripts makes it even more challenging to tackle the OOV problem. So, in this paper, we calculate the vocabulary of each corpus and compare it to show the effect of vocabulary size in training models. We consider the number of unique tokens as vocabulary size.

C. Number of Sentences

Machine translation and conversational systems depend on the millions of parallel sentences to create a better system. So we choose the number of sentences as one of the metric.

978-1-7281-5197-7/20/$31.00 ©2020 IEEE

D. Average Sentence Length

Average sentence length can be used as a measure of grammatical complexity based on the assumption that longer sentence has a more complex syntactic and semantic structure than shorter sentences. It also shows richness and descriptiveness of sentences in the corpus [19]–[21].

TABLE I
DATASETS FOR CODE SWITCHING

NLP Task	Corpora	Languages
Language Identification and POS-Tagging	[22]–[29]	Mandarin-Taiwanese, English-Spanish, Mandarin-English, Nepali-English, Hindi-Nepali, Bengali, Arabic Dialectal-Arabic, Spanish-English, English-Hindi
Named Entity Recognition	[26], [28], [30]–[33]	English-Spanish, English-Egyptian, Modern Standard Arabic-Egyptian, English-Tamil, English-Hindi, Hindi-English
Sentiment Analysis	[26], [34]–[39]	English-Chinese, English-Spanish, English-Hindi, English-Bengali
Conversational Systems	[40]–[42]	n Hindi-English, Bengali-English, Gujarati-English, Tamil-English
Machine Translation	[43]–[45]	English-Hindi, English-Arabic

IV. DATASETS

Despite initial efforts into Code-Switching research started years ago, advancement in the research community is slow. The main difficulties in solving this problem stem from the non-availability of enough data. In this section, we state the different areas of research and list out the available current datasets to that particular research area.

A. Language Identification and POS-Tagging

Language identification at the document level [46]–[48] have been studied extensively as classification problem. More fine-grained language identification at word level is crucial for code-switching corpora to select the right parsers to process multilingual sentences and to build a resource for code-switching corpora.

Code-Switching speech corpus was used to identify the language in each utterances using cues [22] published in INTERSPEECH 2008. Their paper studies the Mandarin-Taiwanese code-switching utterance and proposed a language identification system that integrates acoustics, prosodic and phonetics cues. The data was collected from both Mandarin-speaking and Taiwanese-speaking individuals. The data was not published for public use at that time. At the same time Solorio and Liu from the University of Texas [23], created data by recording a conversation among three English-Spanish speakers for 40 minutes and manually transcribed and annotated with Part-Of-Speech tags. Pos-Tagging for code-switching was first explored for English-Spanish data by Solorio and Liu [23]. We were not able to collect this dataset as well. Romanized and code-switched text data was collected by [24], they collected 1,239 code-mixed posts and the trilingual corpus of 12K Facebook posts contains word-level language annotations and POS tags.

Diab et al [25], [26] organized a shared task that covers four language pairs and is focused on social media data. This data is annotated data from Twitter for Modern Standard Arabic-Arabic dialects, Mandarin-English, Nepali-English, Spanish English for language identification. The data can be downloaded from a shared task website [1]. In 2015, Das et al [27], [28] released the POS tagged dataset for Code mixing at the ICON-2015 NLP tool contest. The task is to identify the POS tags for three Indian Languages (Hindi, Bengali, and Telugu). In the following year, the added more data from Facebook and Whatsapp to ICON 2016 NLP contest. To make a high proportion of code mixed dataset Singh et al [29] released the English-Hindi code-switching dataset.

B. Named Entity Recognition

Named Entity Recognition (NER) is a tagging problem that is challenging for social media data because of inherent noisiness. In addition to improper syntax and semantic structure, the code-switching makes it even more challenging. In 2017, Derczynski organized the shared task on Novel and Emerging Entity Recognition [30] and released the dataset. This dataset contains 1000 annotated tweets, totaling 65,124 tokens. The shared task on [32], the divided the task into two competition based on English-Spanish and English-Egyptian language pairs. They used the language identification dataset from CALCS [26], [28] to annotate entity tags. English-Tamil and English-Hindi NER datasets were created by crawling tweets [31]. Following, Singh et al [33] created Hindi-English code-mixed tweets for NER.

C. Sentiment Analysis

Sentiment analysis has several applications from opinion mining to social analysis, it becomes more useful to analyze code-switched data from social media data for multilingual societies. However, the suitable annotated dataset for this area is not widely available. English-Hindi corpus was created by [38], they collected data from fan pages and hate pages of Virat Kohli [2] from Facebook and Google Plus. they collected 180 code-switched sentences. English-Spanish corpus was introduced by [34], [35] the first code-switching corpus with sentiment labels. The authors collected tweets based on the training collection from [26] shared the task. Chinese-English [39] was collected from Weibo.com. This dataset was

[1] https://www.emnlp2014.org/workshops/CodeSwitch/call.html
[2] famous Indian cricket player at the time of writing

TABLE II
STATISTICS OF CODE SWITCHING DATASETS

Dataset	Language pair	Number of Words or Tokens	Vocabulary Size	Number of Sentences	Average Sentence Length	Paper
Code-Switching shared task	Spanish-English	-	-	11,400	-	[25], [26]
	Nepali-English	-	-	146,055	-	
	Modern Standard Arabic-Arabic dialects	-	-	11,9316	-	
	Mandarin-English	-	-	17,430	-	
Named Entity Recognition	English-Hindi	11,3667	5007	3,638	5.6	[32], [33]
	English-Spanish	825,151	-	67,223	-	
	Modern Standard Arabic-Egyptian	248,478	-	12,334	-	
Sentiment Analysis	English-Hindi	-	-	180	-	[38]
	English-Spanish	-	-	3,062	-	
	Chinese-English	-	-	2,312	-	[39]
	Hindi-English	59,899	7,549	3,879	15	[36]
	Hindi-English	-	-	18,461	-	[37]
	Bengali-English	-	-	5,538	-	[37]
Conversational System	Hindi-English	972528	1,676	6,549	8.16	[40]
	Bengali-English	613,433	1,372	6,274	7.74	
	Gujarati-English	935,232	1,858	6,417	8.04	
	Tamil-English	903,003	2,185	6,666	6.78	
	English-Hindi	-	-	7,700	-	[42]
	English-Hindi	-	-	23,100	-	
Machine Translation	English-Hindi	63,913	-	6,096	-	[43]
	Arabic-English	508,000,000	107,8000	9,700,000		[44]
	English-Hindi	17,920	-	-	-	[45]

annotated for five emotions, namely happiness, sadness, fear, anger, and surprise. The authors have created their annotation tool and format to annotate the data.

The authors [36] collected user comments from public Facebook pages popular in India, like celebrity pages and political leader pages. They manually pre-processed to remove the comments that were written in the native script and longer sentences. They have a 3-level polarity scale. Two code-switching data pairs English-Hindi and English-Bengali were released by Patra et al. [37] which was collected from Twitter.

D. Conversational System

In recent years, there has been increasing popularity with the conversational system as a virtual assistant. Work previous to [40] considered only monolingual corpus. Banerjee et. al, [40] build code-mixed goal-oriented conversations for four Indian languages from the DSTC2 restaurant reservation dataset to mix of code-switched corpora with the help of crowdsourced workers. The data is 5-way parallel code-mix corpora. Following this, Jayarao and Srivastava [41] constructed two datasets for English-Hindi pair with the help of content writes and Google translate. For question answering system, [42] have created a code-switching corpora for three languages- Hindi-English, Telugu-English, and Tamil-English. They followed two modes of data collection, they are collecting code-switched questionf from images and articles.

E. Machine Translation

Chakravarthi et al. [49]–[51] studied the the effects of code-switching in their work to improve the machine translation for under-resourced languages and how the code-switching in the parallel corpora from OPUS project website [3] affected the the quality of translation. Dhar et al. [43] took effort to create machine translation system for code-mixed content. To create machine translation, sentence aligned parallel corpora is required. They created new parallel corpus for English-Hindi code-switched pair. The corpus contains gold standard parallel corpus of 6096 English-Hindi sentences. They were used to improve existing machine translation systems by augmenting the data to the parallel corpora. For English-Arabic, [44] created parallel corpus by extracting UN documents between January 2000 an September 2009. This corpus is multilingual code-switching corpus, containing not only English in Arabic side of corpus but also French, Spanish and other languages. Sign and Solorio [45] created a Hindi-English parallel corpus containing code-switching from Facebook data. The dataset also considers the spelling variants in social media code-switching.

V. ANALYSIS

One of the interesting things we realized after the analysis was most of the dataset collected from Twitter was published

[3]http://opus.nlpl.eu/

only tweet ids. We tried to download the tweets from Twitter based on the tweet ids using Twitter API [4]. The number of tweets that were released for the shared task or with the paper was higher than what we could recollect while writing the paper. Many of the tweets were deleted from Twitter. It made it challenging to do the quality checks using our metric that we defined int the Section III. Many of the datasets showed statistics of the number of tokens in each language used in code-mixed sentences. Not all datasets are accessible easily. While some of the datasets are well-defined and easily available for public use. As seen in Table II, we can see the influence of English with Arabic-English showing dominance with a large amount of data. Some under-resourced languages such as Hindi, Tamil, Bengali, Gujarati also show up in this list. As one can notice not all the data from the datasets mentioned in Table II not all the details of the datasets are available as datasets are not accessible.

VI. Conclusion

Code-switching is an emerging trend in the natural language processing research community and multilingual speakers. The application of code-switching in tasks such as Named Entity Recognition, Sentiment Analysis, and Conversational System increases the value of such datasets. The research community has addressed this trend by introducing challenges on language identification, POS tagging and conducting the shared task. From the challenges and shared tasks, there were many datasets released. Due to privacy issues and other regulations from different social media platform only the ids were published. Not all the data were available to download now. We have shown the available data and their statistics with our quality metrics. From this, we conclude that very few datasets are available for the research in code-switching data.

References

[1] J. Androutsopoulos, "Code-switching in computer-mediated communication," *Pragmatics of computer-mediated communication*, pp. 667–694, 2013.

[2] M. W. Tay, "Code switching and code mixing as a communicative strategy in multilingual discourse," *World Englishes*, vol. 8, no. 3, pp. 407–417, 1989.

[3] C. Nilep, ""code switching" in sociocultural linguistics," *Colorado Research in Linguistics*, vol. 19, no. 1, p. 1, 2006.

[4] C. M. Scotton, "The possibility of code-switching: motivation for maintaining multilingualism," *Anthropological linguistics*, pp. 432–444, 1982.

[5] L. Milroy and W. Li, *A social network approach to code-switching*. Cambridge University Press, 1995.

[6] V. Moodley, "Codeswitching in the multilingual english first language classroom," *International Journal of Bilingual Education and Bilingualism*, vol. 10, no. 6, pp. 707–722, 2007.

[7] C. Meinhof, "Principles of practical orthography for african languages—i." *Africa*, vol. 1, no. 2, pp. 228–236, 1928.

[8] M. Cahill and K. Rice, *Developing orthographies for unwritten languages*. SIL International Dallas, Texas, 2014.

[9] B. Saint-Jacques, "The roman alphabet in the japanese writing system," *Visible Language*, vol. 21, no. 1, p. 88, 1987.

[10] A. Rosowsky, "'writing it in english': script choices among young multilingual muslims in the uk," *Journal of Multilingual and Multicultural Development*, vol. 31, no. 2, pp. 163–179, 2010.

[11] M. Warschauer, G. R. E. Said, and A. G. Zohry, "Language choice online: Globalization and identity in egypt," *Journal of Computer-Mediated Communication*, vol. 7, no. 4, p. JCMC744, 2002.

[12] N. Jain and R. A. Bhat, "Language identification in code-switching scenario," in *Proceedings of the First Workshop on Computational Approaches to Code Switching*. Doha, Qatar: Association for Computational Linguistics, Oct. 2014, pp. 87–93. [Online]. Available: https://www.aclweb.org/anthology/W14-3910

[13] A. Bolonyai, *Code-switching, imperfect acquisition, and attrition*, ser. Cambridge Handbooks in Language and Linguistics. Cambridge University Press, 2009, p. 253–269.

[14] K. Bali, J. Sharma, M. Choudhury, and Y. Vyas, ""I am borrowing ya mixing ?" an analysis of English-Hindi code mixing in Facebook," in *Proceedings of the First Workshop on Computational Approaches to Code Switching*. Doha, Qatar: Association for Computational Linguistics, Oct. 2014, pp. 116–126. [Online]. Available: https://www.aclweb.org/anthology/W14-3914

[15] O. A. Offiong, B. A. Okon *et al.*, "Code switching as a countenance of language interference: The case of the efik bilingual," *International Journal of Asian Social Science*, vol. 3, no. 4, pp. 899–912, 2013.

[16] J. Gumperz, *Discourse Strategies*, ser. Studies in interactional sociolinguistics. Cambridge University Press, 1987. [Online]. Available: https://books.google.co.in/books?id=l7KDngEACAAJ

[17] I. Hamed, M. Elmahdy, and S. Abdennadher, "Collection and analysis of code-switch Egyptian Arabic-English speech corpus," in *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. Miyazaki, Japan: European Language Resources Association (ELRA), May 2018. [Online]. Available: https://www.aclweb.org/anthology/L18-1601

[18] U. Barman, A. Das, J. Wagner, and J. Foster, "Code mixing: A challenge for language identification in the language of social media," in *Proceedings of the First Workshop on Computational Approaches to Code Switching*. Doha, Qatar: Association for Computational Linguistics, Oct. 2014, pp. 13–23. [Online]. Available: https://www.aclweb.org/anthology/W14-3902

[19] Z. Islam, A. Mehler, and R. Rahman, "Text readability classification of textbooks of a low-resource language," in *Proceedings of the 26th Pacific Asia Conference on Language, Information, and Computation*. Bali, Indonesia: Faculty of Computer Science, Universitas Indonesia, Nov. 2012, pp. 545–553. [Online]. Available: https://www.aclweb.org/anthology/Y12-1059

[20] F. Ferraro, N. Mostafazadeh, T.-H. Huang, L. Vanderwende, J. Devlin, M. Galley, and M. Mitchell, "A survey of current datasets for vision and language research," in *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. Lisbon, Portugal: Association for Computational Linguistics, Sep. 2015, pp. 207–213. [Online]. Available: https://www.aclweb.org/anthology/D15-1021

[21] G. Borbély and A. Kornai, "Sentence length," in *Proceedings of the 16th Meeting on the Mathematics of Language*. Toronto, Canada: Association for Computational Linguistics, 18–19 Jul. 2019, pp. 114–125. [Online]. Available: https://www.aclweb.org/anthology/W19-5710

[22] D. Lyu and R. Lyu, "Language identification on code-switching utterances using multiple cues," in *INTERSPEECH 2008, 9th Annual Conference of the International Speech Communication Association, Brisbane, Australia, September 22-26, 2008*, 2008, pp. 711–714. [Online]. Available: http://www.isca-speech.org/archive/interspeech_2008/i08_0711.html

[23] T. Solorio and Y. Liu, "Learning to predict code-switching points," in *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*. Honolulu, Hawaii: Association for Computational Linguistics, Oct. 2008, pp. 973–981. [Online]. Available: https://www.aclweb.org/anthology/D08-1102

[24] B. King and S. Abney, "Labeling the languages of words in mixed-language documents using weakly supervised methods," in *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Atlanta, Georgia: Association for Computational Linguistics, Jun. 2013, pp. 1110–1119. [Online]. Available: https://www.aclweb.org/anthology/N13-1131

[25] M. Diab, J. Hirschberg, P. Fung, and T. Solorio, Eds., *Proceedings of the First Workshop on Computational Approaches to Code Switching*. Doha, Qatar: Association for Computational Linguistics, Oct. 2014. [Online]. Available: https://www.aclweb.org/anthology/W14-3900

[4] https://developer.twitter.com/

[26] T. Solorio, E. Blair, S. Maharjan, S. Bethard, M. Diab, M. Ghoneim, A. Hawwari, F. AlGhamdi, J. Hirschberg, A. Chang, and P. Fung, "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*. Doha, Qatar: Association for Computational Linguistics, Oct. 2014, pp. 62–72. [Online]. Available: https://www.aclweb.org/anthology/W14-3907

[27] A. Jamatia, B. Gambäck, and A. Das, "Part-of-speech tagging for code-mixed English-Hindi twitter and Facebook chat messages," in *Proceedings of the International Conference Recent Advances in Natural Language Processing*. Hissar, Bulgaria: INCOMA Ltd. Shoumen, BULGARIA, Sep. 2015, pp. 239–248. [Online]. Available: https://www.aclweb.org/anthology/R15-1033

[28] G. Molina, F. AlGhamdi, M. Ghoneim, A. Hawwari, N. Rey-Villamizar, M. Diab, and T. Solorio, "Overview for the second shared task on language identification in code-switched data," in *Proceedings of the Second Workshop on Computational Approaches to Code Switching*. Austin, Texas: Association for Computational Linguistics, Nov. 2016, pp. 40–49. [Online]. Available: https://www.aclweb.org/anthology/W16-5805

[29] K. Singh, I. Sen, and P. Kumaraguru, "A twitter corpus for Hindi-English code mixed POS tagging," in *Proceedings of the Sixth International Workshop on Natural Language Processing for Social Media*. Melbourne, Australia: Association for Computational Linguistics, Jul. 2018, pp. 12–17. [Online]. Available: https://www.aclweb.org/anthology/W18-3503

[30] L. Derczynski, E. Nichols, M. van Erp, and N. Limsopatham, "Results of the WNUT2017 shared task on novel and emerging entity recognition," in *Proceedings of the 3rd Workshop on Noisy User-generated Text*. Copenhagen, Denmark: Association for Computational Linguistics, Sep. 2017, pp. 140–147. [Online]. Available: https://www.aclweb.org/anthology/W17-4418

[31] D. K. Gupta, Shweta, S. Tripathi, A. Ekbal, and P. Bhattacharyya, "A hybrid approach for entity extraction in code-mixed social media data," in *Working notes of FIRE 2016 - Forum for Information Retrieval Evaluation, Kolkata, India, December 7-10, 2016*, 2016, pp. 298–303. [Online]. Available: http://ceur-ws.org/Vol-1737/T7-3.pdf

[32] G. Aguilar, F. AlGhamdi, V. Soto, M. Diab, J. Hirschberg, and T. Solorio, "Named entity recognition on code-switched data: Overview of the CALCS 2018 shared task," in *Proceedings of the Third Workshop on Computational Approaches to Linguistic Code-Switching*. Melbourne, Australia: Association for Computational Linguistics, Jul. 2018, pp. 138–147. [Online]. Available: https://www.aclweb.org/anthology/W18-3219

[33] V. Singh, D. Vijay, S. S. Akhtar, and M. Shrivastava, "Named entity recognition for Hindi-English code-mixed social media text," in *Proceedings of the Seventh Named Entities Workshop*. Melbourne, Australia: Association for Computational Linguistics, Jul. 2018, pp. 27–35. [Online]. Available: https://www.aclweb.org/anthology/W18-2405

[34] D. Vilares, M. A. Alonso, and C. Gómez-Rodríguez, "Sentiment analysis on monolingual, multilingual and code-switching twitter corpora," in *Proceedings of the 6th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*. Lisboa, Portugal: Association for Computational Linguistics, Sep. 2015, pp. 2–8. [Online]. Available: https://www.aclweb.org/anthology/W15-2902

[35] D. Vilares, M. A. Alonso, and C. Gómez-Rodríguez, "En-es-cs: An english-spanish code-switching twitter corpus for multilingual sentiment analysis," in *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, N. C. C. Chair), K. Choukri, T. Declerck, S. Goggi, M. Grobelnik, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, and S. Piperidis, Eds. Paris, France: European Language Resources Association (ELRA), may 2016.

[36] A. Joshi, A. Prabhu, M. Shrivastava, and V. Varma, "Towards sub-word level compositions for sentiment analysis of Hindi-English code mixed text," in *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*. Osaka, Japan: The COLING 2016 Organizing Committee, Dec. 2016, pp. 2482–2491. [Online]. Available: https://www.aclweb.org/anthology/C16-1234

[37] B. G. Patra, D. Das, and A. Das, "Sentiment analysis of code-mixed indian languages: An overview of sail_code-mixed shared task@ icon-2017," *arXiv preprint arXiv:1803.06745*, 2018.

[38] D. Sitaram, S. Murthy, D. Ray, D. Sharma, and K. Dhar, "Sentiment analysis of mixed language employing hindi-english code switching," in *2015 International Conference on Machine Learning and Cybernetics (ICMLC)*, vol. 1, July 2015, pp. 271–276.

[39] S. Lee and Z. Wang, "Emotion in code-switching texts: Corpus construction and analysis," in *Proceedings of the Eighth SIGHAN Workshop on Chinese Language Processing*. Beijing, China: Association for Computational Linguistics, Jul. 2015, pp. 91–99. [Online]. Available: https://www.aclweb.org/anthology/W15-3116

[40] S. Banerjee, N. Moghe, S. Arora, and M. M. Khapra, "A dataset for building code-mixed goal oriented conversation systems," in *Proceedings of the 27th International Conference on Computational Linguistics*. Santa Fe, New Mexico, USA: Association for Computational Linguistics, Aug. 2018, pp. 3766–3780. [Online]. Available: https://www.aclweb.org/anthology/C18-1319

[41] P. Jayarao and A. Srivastava, "Intent detection for code-mix utterances in task oriented dialogue systems," *CoRR*, vol. abs/1812.02914, 2018. [Online]. Available: http://arxiv.org/abs/1812.02914

[42] K. Chandu, E. Loginova, V. Gupta, J. v. Genabith, G. Neumann, M. Chinnakotla, E. Nyberg, and A. W. Black, "Code-mixed question answering challenge: Crowd-sourcing data and techniques," in *Proceedings of the Third Workshop on Computational Approaches to Linguistic Code-Switching*. Melbourne, Australia: Association for Computational Linguistics, Jul. 2018, pp. 29–38. [Online]. Available: https://www.aclweb.org/anthology/W18-3204

[43] M. Dhar, V. Kumar, and M. Shrivastava, "Enabling code-mixed translation: Parallel corpus creation and MT augmentation approach," in *Proceedings of the First Workshop on Linguistic Resources for Natural Language Processing*. Santa Fe, New Mexico, USA: Association for Computational Linguistics, Aug. 2018, pp. 131–140. [Online]. Available: https://www.aclweb.org/anthology/W18-3817

[44] M. A. Menacer, D. Langlois, D. Jouvet, D. Fohr, O. Mella, and K. Smaïli, "Machine translation on a parallel code-switched corpus," in *Advances in Artificial Intelligence*, M.-J. Meurs and F. Rudzicz, Eds. Cham: Springer International Publishing, 2019, pp. 426–432.

[45] T. D. Singh and T. Solorio, "Towards translating mixed-code comments from social media," in *Computational Linguistics and Intelligent Text Processing*, A. Gelbukh, Ed. Cham: Springer International Publishing, 2018, pp. 457–468.

[46] W. B. Cavnar and J. M. Trenkle, "N-gram-based text categorization," in *Proceedings of SDAIR-94, 3rd Annual Symposium on Document Analysis and Information Retrieval*, Las Vegas, US, 1994, pp. 161–175.

[47] B. Hughes, T. Baldwin, S. Bird, J. Nicholson, and A. MacKinlay, "Reconsidering language identification for written language resources," in *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*. Genoa, Italy: European Language Resources Association (ELRA), May 2006.

[48] T. Baldwin and M. Lui, "Language identification: The long and the short of the matter," in *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, ser. HLT '10. Stroudsburg, PA, USA: Association for Computational Linguistics, 2010, pp. 229–237. [Online]. Available: http://dl.acm.org/citation.cfm?id=1857999.1858026

[49] B. R. Chakravarthi, M. Arcan, and J. P. McCrae, "Improving wordnets for under-resourced languages using machine translation," p. 78, 2018.

[50] B. R. Chakravarthi, R. Priyadharshini, B. Stearns, A. Jayapal, S. S, M. Arcan, M. Zarrouk, and J. P. McCrae, "Multilingual multimodal machine translation for Dravidian languages utilizing phonetic transcription," in *Proceedings of the 2nd Workshop on Technologies for MT of Low Resource Languages*. Dublin, Ireland: European Association for Machine Translation, 20 Aug. 2019, pp. 56–63. [Online]. Available: https://www.aclweb.org/anthology/W19-6809

[51] B. R. Chakravarthi, M. Arcan, and J. P. McCrae, "WordNet gloss translation for under-resourced languages using multilingual neural machine translation," in *Proceedings of the Second Workshop on Multilingualism at the Intersection of Knowledge Bases and Machine Translation*. Dublin, Ireland: European Association for Machine Translation, 19 Aug. 2019, pp. 1–7. [Online]. Available: https://www.aclweb.org/anthology/W19-7101