A Survey of Code-switched Speech and Language Processing

Sunayana Sitaram

Microsoft Research India

Khyathi Raghavi Chandu, Sai Krishna Rallabandi, Alan W Black

Carnegie Mellon University

Abstract

Code-switching, the alternation of languages within a conversation or utterance, is a common communicative phenomenon that occurs in multilingual communities across the world. This survey reviews computational approaches for code-switched Speech and Natural Language Processing. We motivate why processing code-switched text and speech is essential for building intelligent agents and systems that interact with users in multilingual communities. As code-switching data and resources are scarce, we present a comprehensive list of datasets available in various code-switched language pairs with the language processing tasks they can be used for. We discuss shared tasks and benchmarks that have been proposed to evaluate language processing systems on code-switched text and speech. We review code-switching research in various Speech and NLP applications, including language processing tools and end-to-end systems. We discuss the evaluation of code-switched speech and NLP systems, including recently proposed benchmarks. We conclude with future directions and open problems in the field.

Keywords: code-switching, multilingualism, speech processing, Natural Language Processing, survey

1. Introduction

Linguistic code choice refers to the use of a language for a specific communicative purpose and **code-switching** denotes a shift from one language to another within a single utterance. Not only is there a plethora of different languages across the world, but speakers also often mix these languages within the same utterance. In fact, some form of code-switching is expected to occur in almost every scenario that involves multilinguals [1]. This can go beyond mere insertion of borrowed words, fillers and phrases, and include morphological and grammatical mixing. Such shifts not only convey group identity [2], embody societal patterning [3] and signal cultural discourse strategies [4] but also have been shown to reduce the social and interpersonal distance [5] in both formal [6, 7] and informal settings.

In this paper we refer to this phenomenon as **code-switching**, though the term code-mixing is also used. While such switching is typically considered informal - and is more likely to be found in speech and in casual text as now found in social media - it is also found in semi formal and formal settings such as news paper headlines and teaching. Therefore, we argue that code-switching should not be looked down upon or ignored but be acknowledged as a genuine form of communication that deserves analysis and development of tools and techniques to be handled appropriately. As language technologies improve and permeate more and more applications that involve interactions with humans [8, 9], it is imperative that they take phenomena such as code-switching into account for any consumer facing technology.

Code-switching is most common among peers who have similar fluency in each language. For example fluent bilingual Spanish and English people may often float between their languages, in a form of communication called Spanglish. Indian sub-continent residents, who often have a substantial fluency in English will often mix their speech with their regional languages in Hinglish (Hindi), Tenglish (Telugu), Tamlish (Tamil) and others. But it is not just English that code-switching occurs with. Southern Mainland Chinese residents who, for ex-

ample, speak Cantonese and Shanghaiese, may switch with Putonghua (standard Mandarin). Arabic Dialects are often mixed with Modern Standard Arabic. The distinction between languages and dialects is hard to define, but we see that code-switching appears with dialects too. African American Vernacular English (AAVE) speakers will commonly switch between AAVE and Standard American English; Scottish people may switch between Scots and Standard English. At an extreme, code-switching could also be used to describe register shifting in monolingual speech. Formal speech versus slang or swearing may follow similar functions and patterns as those in code-switching among two distinct languages.

1.1. Why should we care about code-switched language processing?

It is important to realize that humans are good at constructing language registers and learning new communication methods. Not only are we good at doing this with human-human communication, we also construct and learn to use such registers for human-machine communication efficiently, such as Linux command-line expressions, or the grammar of Alexa interactions. If we want machines to partake in such human-human conversations, we need to also be able to understand what is being said in these varied registers.

For the large companies, understanding code-switched communication will enable better advertisement-targeting. Understanding genuine user sentiment about aspects of products helps improve future versions. [10] found a correlation between language use and sentiment, showing that ignoring one language in favor of the other, or ignoring code-switched languages altogether may lead to the wrong conclusions about user sentiment. For healthcare, understanding how people feel, if they are being open, will help to give better care, and enable better communication with patients, and better distribution and uptake of preventative care. For educators, communication in the right register for tutoring, or understanding if concepts are or are not understood is crucial. For entertainment, non-playing characters should communicate in the appropriate register for the game, and/or be able to understand natural code-switched communication with other players.

Unlike pidgins or creoles [11, 12, 13], where speakers may not have full fluency in the language of influence, we are primarily interested in situations where participants have fluency in each of the languages but are choosing not to stay within one language. Code-switching is not a simple linguistic phenomena and depending on the languages involved, and the type of code-switching the interaction between the component languages may be quite different. It is easy to identify at least linguistic sharing, cross-lingual transfer, lexical borrowing as well as speech errors with restarts commonly within the code-switched data. Likewise although there may be language technology tasks that can be achieved with straightforward techniques, it is clear that some tasks, such as semantic role labeling will require complex cross-lingual analysis.

Many have identified the notion of a matrix language in code-switching [14], that there is an underlying language choice which mostly defines the grammar and morphological aspects of the utterance. From a language technologies point of view, especially when considering code-switched data generation using any form of language modeling, it is possible to identify 'bad' code-switching or even 'wrong' code-switching. Although it is obviously not a binary decision, there are extremes that will almost always be wrong. We cannot in general randomly choose which language a word would be realized in, or simply state that we will choose alternate languages for each word. That is, there are constraints, there is an underlying grammar and there are multiple linguistic theories that have been proposed for code-switching. Modeling the grammar is challenging, even if there may be an eventual standardized Hinglish that everyone in Northern India may speak, at present, such code-switched languages are very dynamic, and will have very diverse ideolects across speakers. This is reminiscent of pidgins and creoles which can develop over time, but they too, especially as they are not normally written languages, are also diverse.

But we should not give up, there is underlying structure, and there are constraints, and we have good machine learning modeling techniques that can deal with uncertainty. Recently, there has been quite a lot of interest in the speech and NLP community on processing code-switched speech and text, and

this paper aims at describing progress made in the field, and discussing open problems.

This survey is organized as follows. First, we introduce why code-switching is a challenging and important problem for speech and NLP. Next, in Section 2, we briefly describe linguistic studies on code-switching with other theoretical aspects. In Section 3 we describe speech and NLP corpora and resources that have been created for code-switched language pairs. Section 4 describes techniques for building models for code-switching in specific speech and NLP applications. Section 5 describes various shared tasks and challenges that have been conducted to evaluate code-switching, and introduces benchmarks that evaluate models across tasks and languages. We conclude in Section 6 with a description of the challenges that remain to be addressed and future directions.

2. Background

Research on code-switching is not recent, and this phenomenon has been studied by linguists for decades. In this section, we provide a description of linguistic studies on code-switching and how to characterize code-switched languages. We do not attempt to be comprehensive, since providing a complete description of code-switching research in linguistics is out of the scope of this paper.

2.1. Types of code-switching

Code-switching is defined as the 'juxtaposition within the same speech exchange of passages of speech belonging to two different grammatical systems or sub-systems'[15], while code-mixing is 'the embedding of linguistic units such as phrases, words and morphemes of one language into an utterance of another language'[16]. The distinction between code-switching, mixing and lexical borrowing is often not clear and can be thought of as lying on a continuum [17]. In this paper, we use the terms 'code-switching' and 'code-mixing' interchangeably, although the distinction between the two may be important for certain applications.

The extent and type of code-switching can vary across language pairs. [18] used word-level Language Identification to estimate which language pairs were code-switched on Twitter. They found that around 3.5% of tweets were code-switched, with the most common pairs being English-Spanish, English-French and English-Portuguese. English-German tweets typically had only one switch point, implying that the tweets usually contained translations of the same content in English and German, while English-Turkish tweets had the most switch points, implying fluid switching between the two languages. Code-switching can also vary within a language pair. For example, casual conversational Hinglish may be different from Hinglish used in Bollywood movies, which may be different from Hinglish seen on Twitter.

2.2. Linguistic Models of Code-switching

Early approaches investigate code-switching by laying down a formal framework taking into account the two grammatical systems of the languages being mixed and a mechanism to switch between these two systems at the intra-sentential level [19]. This model mainly explores asymmetric relations between the two grammars, without an explicit formalism of a third grammar and the understanding of where and how to switch closed class items.

Quantitative analysis conducted by [20] revealed two constraints (1) Free Morpheme Constraint and (2) Equivalence Constraint that function simultaneously. The Free Morpheme constraint specifies that it is possible to switch between full sentences as well as any constituent within the sentence if a free morpheme is present in a constituent. The Equivalence Constraint specifies that language switches generally occur at points where there is no violation of syntactic rules of the participating languages.

[21] worked on incorporating both linguistic and extra-linguistic factors into a single analytical model. This study concludes that there are no visibly ungrammatical combinations of the two languages and code-switching is independent of the bilingual ability of the speaker. [22] showed that there exists a constituent tree labeling, implying that around a switch point there is a constraint on an

6

equivalence order in constituents. The above described linguistic theories are also used in [23] to identify governing relationships between constituents. [24] have demonstrated evidence that a constrained Universal Grammar needs refinement of f-selection in code-switching as compared to monolingual speech. [25] have proposed four categories for any switch point comprising of harmonization, neutralization, compromise, and blocking.

[26] have a rather interesting approach towards analyzing grammatical variants in code-switching based on pre-conceptualized assumptions. They claim that grammar in this context is subject to poly-idiolectal repertoires of bilingual speakers and sociolinguistic factors take precedence over grammatical factors. Hence they propose accounting for variability among the bilingual speakers. This same work was extended later to examine intra-sentential switching focusing on bilingual compound verbs and using grammatical knowledge.

The linguistic theories mentioned above were put to use in computational frameworks by [27]. They address several issues such as the absence of literal level translation pairs, sensitivity to minor alignment errors and the under specification of the original models. The human evaluation of generated sentences reveals that the acceptability of code-switching patterns depend not only on socio-linguistic factors but also cognitive factors. This work was later extended in [28] to perform language modeling by leveraging the theories discussed above to generate synthetic code-switched text.

While on one hand, there are studies of formally constructing grammatical representations to understand the nature of code-switching, there is also work that focuses on understanding the psycho-linguistic aspect of this subject pertaining to how and when this occurs. There are studies pertaining to socially determined and pragmatic choices in the developmental perspective of switching in bilingual infants [29].

Another stream of work talks about the factors triggering code-switching that are attributed to 'cognate' or trigger words including proper nouns, cognate content words with good and moderate form overlap, and cognate function words. [30] have studied attested contact-induced changes based on prior

7

linguistic theories regarding the types of structural changes in calques, distributions, frequencies, inventory and stability. [31] explored three different hypothesis and presented empirical evidence for the same. They are the relationships between (i) cognate stimuli and code-switching, (ii) syntactic information and code-switching, (iii) entrainment in a code-switched conversation among bilinguals. The empirical evidence demonstrates that there is a strong correlation between precedence of cognates and code-switching, relationship between POS tags and code-switching and the convergence of the rate of entrianment in code-switching.

[32] present an integrated representation of inter and intra-sentential phenomena as well as spoken and written modalities of code-switching. This is done in order to make better reuse of the minimally available code-switched data by analyzing various global dimensions such as modality, discourse, granularity, social familiarity and social hierarchy. These properties pave way to potentially footprint corpora and functional derivations. [33] present a systematic approach to analyze code-switching in conversations. Patterns of switching are analyzed in multi-party conversations from Hindi movie scripts to establish identity and social contexts.

2.3. Measuring the amount of code-switching

Various metrics have been proposed to measure the amount of code-switching in corpora. The Code-mixing Index (CMI) [34] is an utterance and corpus level metric proposed to measure the amount of code-switching in corpora by using word frequencies. [35] propose M-Index which quantifies the ratio of languages in the corpora based on the Gini coefficient to measure the inequality of the distribution of languages in the corpus. [36] extend this metric to describe the probability of switching within a corpus by summing up the probabilities that there has been a language switch. This metric is termed Integration Index (I-Index) and has values of in the range from 0 (a monolingual text in which no switching occurs) to 1.

[37] and [38] also propose the following metrics: Language Entropy, Span Entropy, Burstiness and Memory. Language Entropy and Span Entropy are the number of bits needed to represent the distribution of language spans. Burstiness quantifies whether the switching has periodic character or occurs in bursts. Memory captures the tendency of consecutive language spans to be positively or negatively autocorrelated.

[39] propose techniques to automatically determine the matrix language of a code-switched utterance. Although the notion of the matrix language is based on the underlying grammar of the sentence, [39] show that the matrix language can be determined by word-count alone as an approximation. [40] characterize languages as being asymmetric or symmetric depending on whether code switching is insertional or alternating, and show that the same grammatical constraints hold in both cases.

3. Data and resources

Over the last few years, significant progress has been made in the fields of Speech Processing and Natural Language Processing mainly owing to the use of large and powerful Machine Learning models such as Deep Neural Networks (DNNs). DNNs typically require large labeled corpora for training, which can be found for a few languages such as US English, Mandarin and Modern Standard Arabic, which are commonly termed as high-resource languages. In the presence of large datasets, models can be trained to achieve high accuracies on tasks such as Automatic Speech Recognition, Machine Translation and Parsing.

However, most languages in the world do not have the necessary data and resources to create models with high enough accuracies to be used in real-world systems. The situation is even more stark for code-switched languages, since considerable care is taken to leave out foreign words while building monolingual resources. So, even if monolingual resources exist for one or more of the languages being mixed, code-switched speech and language resources are very scarce.

However, owing to the recent interest in code-switched speech and language processing, there are some speech and text data sets available for a few language pairs, which we describe next.

3.1. Speech data

Data used for building Automatic Speech Recognition (ASR) and Text to Speech (TTS) systems typically consists of recorded speech and the corresponding transcripts. For ASR systems, the speech may be spontaneous or read, and typically needs to be at least a few thousand hours to build systems that are usable. For TTS systems, a few hours of clean, well recorded speech from a single speaker is typically enough. Below is a list of code-switched data sets available for speech processing.

- SEAME [41] is a corpus of Mandarin-English code-switching by bilinguals in Singapore and Malaysia, with Mandarin being the dominant language in the recordings. It consists of 63 hours of interviews and conversational speech from 97 speakers. [42] describes the updated SEAME corpus with an additional 129 hours of speech.

- The HKUST Mandarin-English Corpus [43] also consists of interviews and conversational speech with 5 hours of transcribed and 15 hours of untranscribed speech.

- The CECOS corpus [44] contains 12 hours of prompted Mandarin-English speech from 77 speakers.

- The OC16-CE80 corpus is a 80 hour Mandarin-English corpus with English words embedded in Mandarin utterances [45].

- The CUMIX Cantonese-English speech corpus [46] contains 17 hours of code-switched speech read by 80 speakers.

- A Mandarin-Taiwanese corpus is described in [47] containing 4000 utterances recorded by 16 speakers.

- [48] present an artificially generated Japanese-English code-switched corpus using a Japanese and English Text-to-speech system from a bilingual speaker. The corpus consists of 280k speech utterances.

- BANGOR-MIAMI [49] is a Spanish-English code-switched corpus consisting of 56 audio recordings and their corresponding transcripts. The recordings consist of informal conversations between two or more speakers, involving a total of 84 speakers.

- A small Spanish-English corpus consisting of 40 minutes of spontaneous speech is described in [50].

- [51] describe an audio and video corpus of elicited code-switched Spanish-English and Hindi-English dialogues. The corpus consists of over 700 calls to an automated agent by workers on Amazon Mechanical Turk, although only a small subset of these have been transcribed.

- The MSR Hindi-English database [52] consists of 50 hours of conversational speech between Hindi-English bilinguals. There are around 500 speakers in the corpus.

- [53] crawled blogs to collect Hindi-English code-switched utterances. 71 speakers recorded around 7000 utterances which were then transcribed and used to build an ASR system.

- [54] describe the creation of a phonetically balanced Hindi-English corpus for code-switched ASR. This corpus contains read speech from 78 speakers, with each speaker having recorded around a minute of speech. The prompts have been collected from news websites and sampled for phonetic coverage.

- [55] collected a small Hindi-English corpus of student interviews. The corpus contains 3 minutes of transcribed speech from 9 speakers.

- [56] collected 1000 hours of Malay-English speech from 208 Chinese, Malay and Indian speakers.

- An Egyptian Arabic-English speech corpus is described in [57]. It consists of 5.3 hours of speech from interviews with 12 participants, of which 4.5 hours of speech has been transcribed.

- The MCSM database (Maghrebian Code-switching in Media) [58] consists of broadcasts from Morocco, Algeria and Tunisia with varying amounts of code-switching between French and Arabic. The FACST corpus (French Arabic Code-switching Triggered)[59] consists of 7.3 hours of French-Algerian Arabic code-switched speech from 20 bilingual speakers.

- [60] describes a corpus of Turkish-German conversational speech consisting of 5 hours of annotated speech data. The corpus is annotated with speech and orthography information, including inter and intra-word switch points.

- [61] describe a corpus of radio broadcasts in Frisian covering a 50 year time span containing code-switching with Dutch. The corpus consists of 18.5 hours of speech annotated with speaker information, dialect and code-switching details and the presence of background noise/music.

- A corpus of code-switched isiZulu-English consisting of transcribed speech from soap operas is described in [62]. It contains around 16 hours of transcribed speech with code-switching boundary annotations. This corpus is extended in [63, 64] to also include 14 hours of English-isiXhosa, English-Setswana, and English-Sesotho code-switched speech.

- For Sepedi-English, two code-switching speech corpora are available [65]. The Sepedi Radio corpus consists of broadcast speech which has been used for analyzing code-switching in this language pair. The Sepedi Prompted Speech Corpus consists of 10 hours of speech from 20 speakers.

- Microsoft and SpeechOcean.com recently released 60 hours of code-switched speech in three language pairs - Tamil-English, Telugu-English and Gujarati-English as part of a shared task on Spoken Language Identification [66].

- Although no code-switched speech databases exist for Speech Synthesis, bilingual TTS databases are available from the same speakers in a number of Indian languages and English [67].

3.2. Text data

In this section, we describe various resources that exist for processing code-switched text. Since the type of data and resources vary greatly with the task at hand, we describe them separately for each task.

3.2.1. Language Identification (LID)

Language ID data sets consist of code-switched sentences that are labeled at the word-level with language information. Conventional LID systems operate at the sentence or document level, which leads to the requirement of word-level LID for code-switched sentences. A couple of shared tasks played an important role in establishing datasets for language identification [68] [69].

- A predetermined set of 11 users of Facebook users were selected to search for publicly available content that resulted in 2335 posts and 9813 comments [70]. Two levels of annotations were performed on this data comprising of different levels of code-switching and language tags including *English, Bengali, Hindi, Mixed, Universal* and *Undefined*. A similar approach is also followed by [71] to collect more data from Bengali, Hindi and English with finer annotation schema catering to named entities and also explicitly annotating suffixes.

- [72] used a two step process by crawling tweets related to 28 hashtags comprising of contexts ranging from sports, movies, religion, politics etc, that resulted in 811981 tweets. Identifying 3577 users from these tags, tweets from these users are crawled in order to gather more mixed language thus resulting in 725173 distinct tweets written in Roman script. These tweets are annotated with *English*, *Hindi* and *Other* tags.

13

- Another very large scale dataset that is is not explicitly targeted at code-switching but contains it is [73] that addresses curating socially representative text by taking into account geographic, social, topical and multilingual diversity. This corpus consists of Tweets from 197 countries in 53 languages.

- In [74] authors employed different approaches to collect data in two different code mixed scenarios. For Nepali - English, they selected 42 twitter users and collected 2000 code switched tweets from each of them. For Spanish English they performed a geographical based search and used 50 tweets from 135 users after filtering based on manual inspection.

3.2.2. Named Entity Recognition (NER)

Named Entity Recognition (NER) datasets for code-switching are similar to LID datasets, with word-level annotations.

- A shared task was organized to address NER for code-switched texts using around 50k Spanish-English and around 10k Arabic-English annotated tweets [75].

- Twitter is a commonly used source of code-switched data. [76] annotated 3,638 tweets with three Named Entity tags 'Person', 'Organization' and 'Location' using the BIO scheme.

3.2.3. Part of Speech (POS) Tagging

POS tagging data sets consist of code-switched sentences tagged at the word level with POS information.

- Public pages from Facebook pages of three celebrities and the BBC Hindi news page are used to gather 6,983 posts and comments and annotated with POS tags in addition to matrix language information [77].

- ICON 2015 conducted a shared task on POS tagging for which they released data in Hindi-English, Bengali-English, Tamil-English [78]. The dataset contains 1k-3k annotated utterances for each language pair.

- Code-switched Turkish-German tweets were annotated based on Universal Dependencies POS tags and the authors proposed guidelines for the Turkish parts to adopt language-general heuristics to gather a corpus of 1029 tweets [79].

- 922 sentences of spoken Spanish-English conversational data is transcribed and annotated with POS tags in [80].

- [81] gathered 1106 messages (552 Facebook posts and 554 tweets) in Hindi-English and annotated them with a Twitter specific tagset. [82] describe an English-Bengali corpus consisting of Twitter messages and two English-Hindi corpora consisting of Twitter and Facebook messages tagged with coarse and fine grained POS tags.

- [83] crowd-sourced POS tags using the Universal POS tagset to annotate the BANGOR-MIAMI corpus which is a conversational speech dataset with Spanish-English code-switching.

- [84] create a corpus of 886,252 tokens in Modern Standard Arabic-Egyptian Arabic annotated with POS as well as 16 code-switching tags. The code-switching tags include language ID information as well as special tags to label mixed words that contain morphemes from different languages. This dataset consists of tweets and sentences from news, commentaries and discussion forms. The annotation is being extended to other Arabic dialects such as Levantine, Iraqi, Gulf, Moroccan and Tunisian.

3.2.4. Parsing

Datasets for parsing contain code-switched sentences with dependency parses and chunking tags.

- [85] have worked on using monolingual resources to parse low resource languages in the presence of code-switching. While the training data comprises of Russian UD v2.0 corpus with 3,850 sentences and 40 Komi sen-

15

tences, the test set comprises of 80 Komi-Russian multilingual sentences and 25 Komi spoken sentences.

- [86] have presented a dataset of 450 Hindi-English CM tweets for evaluation purposes annotated with dependency parse relations.

- A shallow parsing dataset comprising of 8450 tweets annotated with language id, normalized script, POS tagset and chunking tags is described in [87].

- Code-switched test utterances for the NLmaps corpus are constructed by [88]. They use a parallel corpus of English and German utterances which share the same logical form to construct code-switched utterances. They use 1500 pairs of sentences from each language for training and 880 pairs for testing.

3.2.5. Question Answering (QA)

Question answering (QA) datasets typically consist of questions and answers that are in the form of articles, images, tuples etc. In case of code-switched QA, the questions are typically in code-switched form.

- A first step towards creating a code-switched QA dataset was attempted by collecting 3000 questions from a version of a TV show *"Who wants to be a Millionaire?"* and general knowledge questions from primary school textbooks for Hindi-English code-switching questions [89]. Out of the 3000 questions, 1000 unique questions are used in order to avoid any individual biases of language usage.

- In lieu of addressing the possibility of lexical bias from entrainment in [89], another effort was made on a larger scale to collect 5933 questions for Hindi-English, Tamil-English, Telugu-English grounded on articles and images [90]. The final dataset included 1,694 Hindi-English, 2,848 Tamlil-English and 1,391 Telugu-English factoid questions and their answers [91].

16

- Another section of efforts that move towards using monolingual data from English and weakly supervised and imperfect bilingual embeddings provided a test set of 250 Hindi-English code-switched questions mapped between SimpleQuestions dataset and Freebase tuples [92].

- One of the early efforts also include leveraging around 300 messages from social media platforms like Twitter and blogs to collect 506 questions from the domains of sports and tourism [93].

3.3. Natural Language Inference

[94] present the first dataset for code-switched NLI, in which premises are taken from Bollywood (Hindi) movie scripts and annotators create hypotheses that entail or contradict the premises. The dataset contains 400 premises and around 2k hypotheses.

3.4. Social media datasets

Various datasets from social media such as Facebook and Twitter have been collected for different NLP tasks, which we describe in this section.

[95] collected 1959 Hindi-English tweets and asked annotators to rank tweets according to relevance for specific queries.

[96] collect a Twitter corpus of around 4k tweets and annotate it for Hate Speech. [97] create a corpus of around 3k tweets for automated irony detection.

4. Code-switched Speech and NLP Techniques

[98] present a brief survey of code-switching studies in NLP. [99] describe the challenges in computational processing of core NLP tasks as well as downstream applications. They highlight issues caused due to combining two languages at the lexical and syntactic level, using examples from several tasks and language pairs.

In this paper, we provide a comprehensive description of work done in code-switched speech and NLP. Various approaches have been taken to build speech

17

and NLP systems for code-switched languages depending on the availability of monolingual, bilingual and code-switched data. When there is a complete lack of code-switched data and resources, a few attempts have been made to build models using only monolingual resources from the two languages being mixed.

Domain adaptation or transfer learning techniques can be used, wherein models are built on monolingual data and resources in the two languages and a small amount of 'in-domain' code-switched data can be used to tune the models.

Word embeddings have been used recently for a wide variety of NLP tasks. Code-switched embeddings can be created using code-switched corpora [100], however, in practice such resources are not available and other techniques such as synthesizing code-switched data for training such embeddings can be used. Massive multilingual models such as multilingual BERT [101] have also been explored in code-switched NLP.

4.1. Automatic Speech Recognition

Since code-switching is a spoken language phenomenon, it is important that Automatic Speech Recognizers (ASRs) that are deployed in multilingual communities are able to handle code-switching. In addition, ASR systems tend to be the first step in a pipeline of different systems in applications such as conversational agents, so any errors made by ASR systems can propagate through the system and lead to failures in interactions.

Attempts have been made to approach the problem of code-switched ASR from the acoustic, language and pronunciation modeling perspectives.

Initial attempts at handling code-switched speech recognition identified the language being spoken by using a Language Identification (LID) system and then used the appropriate monolingual decoder for recognition. One approach is to identify the language boundaries and subsequently use an monolingual ASR system to recognize monolingual fragments [102]. Another approach runs multiple recognizers in parallel with an LID system and uses scores from all the systems for decoding speech [103]. In [56], no LID system is used - instead, two recognizers in English and Malay are run in parallel and the hypotheses pro-

duced are re-scored to get the final code-switched recognition result. However, the disadvantages with multi-pass approaches are that errors made by the LID system are not possible to recover from. [47] suggest a single-pass approach with soft decisions on LID and language boundary detection for Mandarin-Taiwanese ASR.

The choice of phone set is important in building ASR systems and for code-switched language pairs, the choice of phoneset is not always obvious, since one language can have an influence on the pronunciation of the other language. [104] develop a cross-lingual phonetic Acoustic Model for Cantonese-English speech, with the phone set designed based on linguistic knowledge. [105] present three approaches for Mandarin-English ASR - combining the two phone inventories, using IPA mappings to construct a bilingual phone set and clustering phones by using the Bhattacharyya distance and acoustic likelihood. The clustering approach outperforms the IPA-based mapping and is comparable to the combination of the phone inventories. [52] describe approaches to combine phone sets, merge phones manually using knowledge and iterative merging using ASR errors on Hindi-English speech. Although the automatic approach is promising, manual merging using expert knowledge from a bilingual speaker performs best. [106] use IPA, Bhattacharya distance and discriminative training to combine phone sets for Mandarin-English. When code-switching occurs between closely related languages, the phone set of one language can be extended to cover the other, as is suggested in [107] for Ukranian-Russian ASR. In this work, the Ukranian phone set and lexicon are extended to cover Russian words using phonetic knowledge about both languages.

[65] describe an ASR system for Sepedi-English in which a single Sepedi lexicon is used for decoding. English pronunciations in terms of the Sepedi phone set are obtained by phone-decoding English words with the Sepedi ASR. [54] use a common Wx-based phone set for Hindi-English ASR built using a large amount of monolingual Hindi data with a small amount of code-switched Hindi-English data. [108] use cross-lingual data sharing to tackle the problem of highly imbalanced Mandarin-English code-switching, where the speakers speak

primarily in Mandarin. In [109], authors attempt to alleviate the problem of L2 word pronunciation by creating linguistically motivated pairwise mappings

When data from both languages is available but there is no or very little data in code-switched form, bilingual models can be built. [106] train the Acoustic Model on bilingual data, while [110] and [111] use existing monolingual models and with a phone-mapped lexicon and modified Language Model for Hindi-English ASR. In [112], authors create synthetic code mixed speech by concatenating segments from different monolingual utterances and employ this to improve Hindi-English code mixed ASR performance. In [113], authors first detect real and untranscribed code mixed segments from online archives. They then employ semi supervised and active learning techniques to obtain transcriptions and use as augmented data to train code switched models. In [114] authors follow semi supervised training and show that incorporating language and speaker information is helpful while building bilingual acoustic models. In [115] authors combine monolingual and bilingual graphs together with a unified acoustic model.

[116] propose a technique known as meta transfer learning to select the best monolingual data for transfer that can improve code-switched models. [117] describe the importance of data selection between subsets of English, Mandarin and code-switched datasets for improving Mandarin-English ASR and show that simply pooling all the data leads to worse results.

[118] build a bilingual DNN-based ASR system for Frisian-Dutch broadcast speech using both language-dependent and independent phones. The language dependent approach, where each phone is tagged with the language and modeled separately performs better. [119] decode untranscribed data with this ASR system and add the decoded speech to ASR training data after rescoring using Language Models. In [120], this ASR is significantly improved with augmented textual and acoustic data by adding more monolingual data in Dutch, automatically transcribing untranscribed data, generating code-switched data using Recurrent LMs and machine translation.

[64, 121] build a unified ASR system for five South African languages, by us-

ing interpolated language models from English-isiZulu, English-isiXhosa, English-Setswana and English-Sesotho. This system is capable of recognizing code-switched speech in any of the five language combinations.

[122] use semi-supervised techniques to improve the lexicon, acoustic model and language model of English-Mandarin code-switched ASR. They modify the lexicon to deal with accents and treat utterances that the ASR system performs poorly on as unsupervised data. In [123] authors utilise ASR and TTS in a semi supervised fashion to learn code switching. They further show that integrating language embeddings allows the framework to address even the language pairs not seen during training [124].

In [125] authors jointly train two Mandarin-English acoustic models that differ in the choice of acoustic units describing the salient acoustic and phonetic information. In [126], they observe that sharing parameters between the primary and auxiliary tasks helps capture language switching information.

Recent studies have explored end-to-end ASR for code-switching. Traditional end-to-end ASR models require a large amount of training data, which is difficult to find for code-switched speech. [127] propose a CTC-based model for Mandarin-English speech, in which the model is first trained using monolingual data and then fine-tuned on code-switched data. [128] use transfer learning from monolingual models, wordpieces as opposed to graphemes and multitask learning with language identification as an additional task for Mandarin-English end-to-end ASR. In [129], authors address the scenario where monolingual speakers attempt to comprehend code switched speech in the context of a dialog. To address this, they build a system to recognize code mixed speech and translate it to monolingual text. In [130], authors present a hypothesis that the discrepancy between distributions of token representations for different languages restricts end to end models. To alleviate this, they constrain the token representations using Shannon divergence and cosine distance. In [131], authors perform a frame level language detection and adjust the posterior distribution with CTC conditioned on the language detection. [132] present an RNN-T model with language bias that can improve upon an RNN-T model without any LID information,

21

without needing an explicit LID system.

In [133], authors explore using two types of units: characters for both Mandarin and English, characters for Mandarin and sub word units for English. In [134] authors employ BPE sub word units. In [135], authors employ a frame level language recognition system to seed CTC based acoustic model.

Recent work by [136] showed that speech recognition models fine-tuned on code-switched data regress on monolingual speech. To alleviate this issue and build robust models that can improve on both monolingual and code-switched speech recognition, the authors propose using Learning Without Forgetting and adversarial training. [137] extends this work by proposing a multi-task approach to domain adversarial training that shows further improvements on both monolingual and code-switched ASR.

As stated earlier, switching/mixing and borrowing are not always clearly distinguishable. Due to this, the transcription of code-switched and borrowed words is often not standardized, and can lead to the presence of words being cross-transcribed in both languages. [138] automatically identify and disambiguate homophones in code-switched data to improve recognition of code-switched Hindi-English speech.

4.2. Language Modeling

Language models (LMs) are used in a variety of Speech and NLP systems, most notably in ASR and Machine Translation. Although there is significantly more code-switched text data compared to speech data in the form of informal conversational data such as on Twitter, Facebook and Internet forums, robust language models typically require millions of sentences to build. Code-switched text data found on the Internet may not follow exactly the same patterns as code-switched speech. This makes building LMs for code-switched languages challenging.

Monolingual data in the languages being mixed may be available, and some approaches use only monolingual data in the languages being mixed [139] while

others use large amounts of monolingual data with a small amount of code-switched data.

Other approaches have used grammatical constraints imposed by theories of code-switching to constrain search paths in language models built using artificially generated data. [140] use inversion constraints to predict CS points and integrate this prediction into the ASR decoding process. [141] integrate Functional Head constraints (FHC) for code-switching into the Language Model for Mandarin-English speech recognition. This work uses parsing techniques to restrict the lattice paths during decoding of speech to those permissible under the FHC theory. [142] assign weights to parallel sentences to build a code-switched translation model that is used with a language model for decoding code-switched Mandarin-English speech.

[143] show that a training curriculum where an Recurrent Neural Network (RNN) LM is trained rst with interleaved monolingual data in both languages followed by code-switched data gives the best results for English-Spanish LM. [100] extend this work by using grammatical models of code-switching to generate artificial code-switched data and using a small amount of real code-switched data to sample from the artificially generated data to build Language Models.

[144] uses Factored Language Models for rescoring n-best lists during ASR decoding. The factors used include POS tags, code-switching point probability and LID. In [145], [146] and [147], RNNLMs are combined with n-gram based models, or converted to backoff models, giving improvements in perplexity and mixed error rate.

[148, 149] investigate the importance of syntactic information such as Part-of-Speech(POS) in predicting the switching point. They observe that the switchig attitude is speaker dependent[150].

[151] synthesize isiZulu-English bigrams using word embeddings and use them to augment training data for LMs, which leads to a reduction in perplexity when tested on a corpus of soap opera speech.

In [152] authors employ dual RNNs for language model training while [153, 154, 155] investigate the applicability of artificially generated code mixed data

for data augmentation.

In [156], authors show that encoding language information improves language model by learning code switch points. In [157] authors present a discriminative training based approach for model code mixed text. Alternatively, authors in [158] propose to manipulate n gram based language model by employing clustering for the infrequent words. In [159] authors present an approach using multi task learning by jointly learning language modeling as well POS tagging.

[160] use a bilingual attention language model that learns cross-lingual probabilities by using parallel data simultaneously along with the language modeling objective and achieves high reductions in perplexity over the SEAME corpus.

4.3. Code-switching detection from speech

In [161] authors show that humans exploit prosodic cues to detect code mixing. They also show taht humans can anticipate switch poitns even in noisy speech.

As mentioned earlier, some ASR systems first try to detect the language being spoken and then use the appropriate model to decode speech. In case of intra-sentential switching, it may be useful to be able to detect the code-switching style of a particular utterance, and be able to adapt to that style through specialized language models or other adaptation techniques.

[162] look at the problem of language detection from code-switched speech and classify code-switched corpora by code-switching style and show that features extracted from acoustics alone can distinguish between different kinds of code-switching in a single language.

In [163, 164] authors investigate the effectiveness of using retrained multilingual DNNS and augmenting the data for detecting the language. In [165, 166] authors employ word based lexical information [167] build HMM based acoustic model followed by an SVM based decision classifier to identify the code mixing between Northern Sotho and English.

4.4. Speech Synthesis

Most Text to Speech (TTS) systems assume that the input is in a single language and that it is written in native script. However, due to the rise in globalization, phenomena such as code-switching are now seen in various types of text ranging from news articles through comments/posts on social media, leading to co-existence of multiple languages in the same sentence. Incidentally, these typically are the scenarios where TTS systems are widely deployed as speech interfaces and therefore these systems should be able to handle such input. Even though independent monolingual synthesizers today are of very high quality, they are not fully capable of effectively handling such mixed content that they encounter when deployed. These synthesizers in such cases speak out the wrong/accented version at best or completely leave the words from the other language out at worst. Considering that the words from other language(s) used in such contexts are often the most important content in the message, these systems need to be able to handle this scenario better.

Current approaches handling code-switching fall into three broad categories: phone mapping, multilingual or polyglot synthesis. In phone mapping, the phones of the foreign language are substituted with the closest sounding phones of the primary language, often resulting in strongly accented speech. In a multilingual setting, each text portion in a different language is synthesised by a corresponding monolingual TTS system. This typically means that the different languages will have different voices unless each of the voices is trained on the voice of same multilingual speaker. Even if we have access to bilingual databases, care needs to be taken to ensure that the recording conditions of the two databases are very similar. The polyglot solution refers to the case where a single system is trained using data from a multilingual speaker. Similar approaches to dealing with code-switching have been focused on assimilation at the linguistic level, and advocate applying a foreign linguistic model to a monolingual TTS system. The linguistic model might include text analysis and normalisation, a G2P module and a mapping between the phone set of the foreign language and the primary language of the TTS system [168, 169, 170].

Other approaches utilise cross-language voice conversion techniques [171] and adaptation on a combination of data from multiple languages[172]. Assimilation at the linguistic level is fairly successful for phonetically similar languages [170], and the resulting foreign synthesized speech was found to be more intelligible compared to an unmodified non-native monolingual system but still retains a degree of accent of the primary language. This might in part be attributed to the non-exact correspondence between individual phone sets.

[173] find from subjective experiments that listeners have a strong preference for cross-lingual systems with Hindi as the target language. However, in practice, this method results in a strong foreign accent while synthesizing the English words. [174, 175] propose a method to use a word to phone mapping instead, where an English word is statistically mapped to Indian language phones.

[176] train speech synthesizers for Hindi-English, Tamil-English and Hindi-Tamil by randomizing the order of bilingual training data which are then used to synthesize monolingual and code-switched text. This leads to improvements in subjective metrics for the code-switched speech and marginal degradation in monolingual speech.

[177] present an end-to-end code-switched TTS for Mandarin English, in which they use bilingual data with a shared encoder that contains language information and separate decoders. [178] extend this approach to use a bilingual phonetic posteriorgram (PPG) to synthesize code-switched speech using only monolingual data. [179] also use a language specific encoder along with a multi-head attention mechanism in the decoder resuling in large improvements in the SEAME corpus.

4.5. Language Identification

The task of lexical level language identification (LID) is one of the skeletal tasks for the lexical level modeling of downstream NLP tasks. Most research has focused on word-level LID, although some work on utterance-level LID also exists. [180] build tools for web-scale analysis of code-switching, using an utterance-level language identification system based on the language ratio of

the two languages involved. A large amount of research in this area has been conducted due to shared tasks on word-level LID ([68], [69]).

Social media data, especially posts from Facebook was used to collect data for the task of LID [70] of Bengali, Hindi and English code-switching. Techniques include dictionary based lookup, supervised techniques applied at word level along with ablation studies of contextual cues and CRF based sequence labeling approaches. Character level n-gram features and contextual information are found to be useful as features.

[181] is among the first computational approaches towards determining intra-word switching by segmenting the words into smaller meaningful units through morphological segmentation and then performing language identification probabilistically. This was followed by intra-word approaches [182, 183] and approaches that incorporate information beyond word level [184, 185, 186, 187, 188, 189, 190, 191, 192, 193]. In addition to features, model based variants have been proposed by [194, 195, 196, 197, 198].

[199] make use of patterns in language usage of Hinglish along with the consecutive POS tags for LID. [71] have also experimented with n-gram modeling with pruning and SVM based models with feature ablations Hindi-English and Bengali-English LID. [72] have worked on re-defining and re-annotating language tags from social media cues based on cultural, core and therapeutic borrowings. [73] have introduced a socially equitable LID system known as EQUILID by explicitly modeling switching with character level sequence to sequence models to encompass dialectal variability in addition to code-switching. [200] present a weakly supervised approach with a CRF based on a generalization expectation criteria that outperformed HMM, Maximum Entropy and Naive Bayes methods by considering this a sequence labeling task.

Recently, POS tagging has also been examined as a means to perform language identification in code-switched scenarios [201]. To this end, they have collected a Devanagari corpus and annotated it with POS tags followed by transliterating it into Roman text. The complementary English data is annotated with POS tags as well. Several classical approaches including SVM, Decision Trees,

27

Logistic Regression and Random Forests have been experimented with. The feature set that included POS tags along with the word length and the word itself with a random forest resulted in the highest performance. Hence monolingual data with corresponding POS tags seem useful in performing language identification of code-switched text.

4.6. Named Entity Recognition

Another sequence labeling task of interest is Named Entity Recognition (NER). [75] organized a shared task on NER in code-switching by collecting data from tweets for Spanish-English and Arabic-English. [202] augmented state-of-the-art character level Convolutional Neural Networks (CNNs) with Bi-LSTMs followed by a CRF layer, by enriching resources from external sources by stacking layers of pre-trained embeddings, Brown clusters and gazetteer lists. [203] attempted to build models from observations from data comprising of less than 3% of surface level Named Entities and a high Out of Vocabulary (OOV) percentage. To address these issues they rely on character based BiLSTM models and leveraging external resources. Prior to this shared task, [204] posed this task as a multi-task learning problem by using a character level CNNs to model non-standard spelling variations followed by a word level Bi-LSTM to model sequences. This work also highlights the importance of gazetteer lists since it is similar to a low resource setting.

[205] studied Arabic text on social media by exploring the influence of word embedding based representations on NER. Along similar lines, [206] also investigated how word representations are capable of boosting semi-supervised approaches to NER. [76] collected tweets from topics like politics, social events, sports and annotated them with three Named Entity Tags in the BIO scheme and explored CRF, LSTM and Decision Tree methods. Formal and informal language specific features were leveraged to employ Conditional Random Fields, Margin Infused Relaxed Algorithm, Support Vector Machines and Maximum Entropy Markov Models to perform NER on informal text in Twitter [207]. [208] collected a dataset in an attempt to benchmark for the task of Named En-

tity Recognition in Arabish from three different sources: Twitter, transcribed conversational speech and translating a standard NER dataset. The dataset comprises of 6k sentences with 130k tokens. The baseline model itself is a BiLSTM-CRF which is one of the heavily investigated architectures in the task of NER. On top of this, they have adopted the FLAIR framework [209] to investigate different types of embeddings along with pooled datasets. They have also experimented with word embeddings that are not only traditionally used and also more recently used such as contextual embeddings in their architecture and discovered that a combination of both performed better. [210] extend the LSTM architecture to combat high percentage of out of vocabulary words in code-switched data using transfer learning with bilingual character representations. Additionally, they also remove the noise with normalization of the spellings. Alternative to the fusion approach see above, [211] utilize the self attention mechanism over the charcacter based embeddings. The final embedding representation is obtained by feeding these word and character based embeddings through a stacked BiLSTM with residual connections. Inspired by this, [212] proposed multilingual meta embeddings that extend the scope to other related and similar languages. They circumvent the problem of lexical level language identification using the same self attention mechanism on pre-trained word embeddings. [213] propose the use of hierarchical meta-embeddings that combine word and sub-word level embeddings to achieve SOTA performance on English-Spanish NER.

4.7. POS Tagging

Recently there has been interest in code-switched structured prediction tasks like POS tagging and parsing. [77] used a dual mechanism of utilizing both a CRF++ based tagger and a Twitter POS tagger in order to tag sequences of mixed language. The same work also proposed a dataset that is obtained from Facebook that is annotated at a multi-level for the tasks of LID, text normalization, back transliteration and POS tagging. They claim that joint modeling of all these tasks is expected to yield better results. [79] presented POS annota-

tion for Turkish-German tweets that align with existing language identification based on POS tags from Universal Dependencies. [80] explored the exploitation of monolingual resources such as taggers (for Spanish and English data) and heuristic based approaches in conjunction with machine learning techniques such as SVM, Logit Boost, Naive Bayes and J48. This work shows that many errors occur in the presence of intra-sentential switching thus establishing the complexity of the task.

[81] have also gathered data from social media platforms such as Facebook and Twitter and have annotated them at coarse and fine grained levels. They focus on comparing language specific taggers with ML based approaches including CRFs, Sequential Minimal Optimization, Naive Bayes and Random Forests and observ that Random Forests performed the best, although only marginally better than combinations of individual language taggers. [83] use crowd-sourcing for annotating universal POS labels for Spanish-English speech data by splitting the task into three subtasks. These are 1. labeling a subset of tokens automatically 2. disambiguating a subset of high frequency words 3. crowd-sourcing tags by decisions based on questions in the form of a decision tree structure. The choice of mode of tagging is based on a curated list of words.

[214] use a stacked model technique and compare them to joint modeling and pipeline based techniques to find that that best stacked model that utilizes all features outperform the joint and pipeline-based models.

[215] carry out normalization of code-switched data and assess the impact on POS tagging as a downstream task. They find that automatic normalization leads to a performance gain in POS tagging.

4.8. Parsing

[216] worked on bilingual syntactic parsing techniques for Hindi-English code-switching using head-driven phrase structure grammar. The parses in cases of ambiguities are ordered based on ontological derivations from Word-Net through a Word Sense Disambiguator [217]. However, there is an assumed external constraint in this work, where the head of the phrase determines the

syntactic properties of the subcategorized elements irrespective of the languages to which these words belong.

[86] leveraged a non-linear neural approach for the task of predicting the transitions for the parser configurations of arc-eager transitions by leveraging only monolingual annotated data by including lexical features from pre-trained word representations. [87] also worked on a pipeline and annotating data for shallow parsing by labeling three individual sequence labeling tasks based on labels, boundaries and combination tasks where a CRF is trained for each of these tasks.

[88] performed multilingual semantic parsing using a transfer learning approach for code-switched text utilizing cross lingual word embeddings in a sequence to sequence framework. [85] compared different systems for dependency parsing and concluded that the Multilingual BIST parser is able to parse code-switched data relatively well.

[218] present a Universal Dependencies dataset in Hindi-English and a neural stacking model for parsing with a new decoding scheme that outperforms prior approaches.

4.9. Question Answering

So far, we have seen individual speech and NLP applications which can be used as part of other downstream applications. One very impactful downstream application of casual and free mixing beyond mere borrowing in terms of information need is Question Answering (QA). This is especially important in the domains of health and technology where there is a rapid change in vocabulary thereby resulting in rapid variations of usage with mixed languages. One of the initial efforts in eliciting code-mixed data to perform question classification was undertaken by [89]. This work leveraged monolingual English questions from websites for school level science and maths, and from Indian version of the show *'Who wants to be a Millionaire?'*. Crowd-workers are asked to translate these questions into mixed language in terms of how they would frame this question to a friend next to them.

Lexical level language identification, transliteration, translation and adjacency features are used to build an SVM based Question Classification model for data annotated based on coarse grained ontology proposed by [219]. Since this mode of data collection has the advantage of gathering parallel corpus of English questions with their corresponding code-switched questions, there is a possibility of lexical bias due to entrainment. In order to combat this, [90] discussed techniques to crowd-source code-mixed questions based on a couple of sources comprising of code-mixed blog articles and based on certain fulcrum images. They organized the first edition of the code-mixed question answering challenge where the participants used techniques based on Deep Semantic Similarity model for retrieval and pre-trained DrQA model fine-tuned on the training dataset. An end-to-end web based QA system WebShodh is built and hosted by [220] which also has an additional advantage of collecting more data.

[92] trained TripletSiamese-Hybrid CNN to re-rank candidate answers that are trained on the SimpleQuestions dataset in monolingual English as well as with loosely translated code-mixed questions in English thereby eliminating the need to actually perform full fledged translation to answer queries. [93] gathered a QA dataset from Facebook messages for Bengali-English CM domain. In addition to this line of work, there were efforts for developing a cross-lingual QA system where questions are asked in one language (English) and the answer is provided in English but the candidate answers are searched in Hindi newspapers [221].

[222] presented a query oriented multi-document summarization system for Telugu-English with a dictionary based approach for cross language query expansion using bilingual lexical resources. Cross language QA systems are explored in European languages as well [223], [224].

4.10. Sentiment Analysis/stance detection

[225] provide a benchmarking dataset to perform sentiment analysis on 3,062 English-Spanish tweets. The annotations are based on SentiStrength into the labels of positive, negative and neutral classes. The same group later extended

this work to compare code-switching in monolingual and multilingual settings [226, 227]. The comparisons made between a multilingual model trained on a multilingual dataset, separate monolingual models, and a monolingual model that is triggered based on the language identification demonstrate the effectiveness of the multilingual model to deal with code-switched scenarios. [228] use a more classical word probabilities based approach to determine the sentiment of a tweet about a movie. In specific, this is performed for tweets in Telugu-English mixed data by transliterating each Roman word to the corresponding Telugu script and computing the probability of the word in each class. [229] conducted a shared task for sentiment analysis of social media data in two language pairs. The dataset used for the shared task includes around 12k and 2500 tweets released for training in Hindi-English and Bengali-English respectively. The best performing system of the shared task used word and character level n-gram features with an SVM classifier. A similar trend is observed by [230] while comparing the models of Naive Bayes and SVM to perform sentiment classification on movie reviews in Bengali-English.

Contrary to this, [231] use CNNs to model sub-word level representations. These are then given to a dual encoder, which both capture the sentiment at the sentence level and at the sub-word level. Similarly, [232] approached this using multitask learning over a CNN based encoder to classify the stance taken on a popular issue of 'Demonetization'. The auxiliary task is posed as a manipulation of the primary task by combining the labels.

Extending this to emotion detection, there have been several attempts to model this as a graph problem. [233] present a proposition of scheme to annotate data collected with emotions for Chinese-English corpus specifically. The schema is developed to address the choice of text in which a sentiment is expressed. This can be either in one of Chinese, or English text, or using both the languages, or using a mixed language text. [234] gathered a dataset from *Weibo.com* which have labelled emotions which are then self aligned among the languages using statistical machine translation paradigm. They use label propagation over a bipartite graph constructed on bilingual and sentiment in-

formation. They have extended this work to joint factor graph model [235] between the two kinds of information identifying the necessity of correlating different emotions as well in addition to sentiment and languages. Along very similar lines, [236] use belief propagation over the factor graphs which poses this as a dynamic programming approach to query a graphical model. The graph itself is constructed as joint factor graph model by utilizing both the bilingual word information and the emotion related information.

4.11. Hate Speech Detection

In [237], authors employ transfer learning. They first train a CNN based model on a large corpus of hateful tweets as source task followed by fine tuning on a transliterated set in the same language. In [238], authors use a combination of psycho-linguistic feature and basic features and perform model averaging. In [239], authors investigate both hierarchical employing phonemic units and sub world level models to detect hate speech from code mixed data.

4.12. Natural Language Inference

[94] present the first work on code-switched NLI, where the task is to predict if a hypotheses entails or contradicts the given premise, which is in the form of a conversation taken from Bollywood (Hindi) movies. They fine-tune multilingual BERT for the task, however, the accuracy of this model is only slightly better than chance showing that NLI is a very challenging problem for code-switched NLP.

4.13. Machine Translation

[240] developed a machine translation scheme for translating Hinglish into pure English and pure Hindi forms by performing cross morphological analysis. [241] show that a zero shot Neural Machine Translation system can also deal with code-switched inputs, however, the results are not as good as monolingual inputs.

4.14. Dialogue and discourse

[242] study lexical and prosodic features of code-switched Hindi-English dialogue and find that the embedded language (English) fragments are spoken more slowly and with more vocal effort, and pitch variation is higher in the code-switched portion of the dialogues compared to the monolingual parts.

[243] treat code-choice as linguistic style and study accommodation across turns in dialogues in Spanish-English and Hindi-English. They find that accommodation is affected by the markedness of the languages in context and is sometimes seen after a few turns, leading to delayed accommodation.

In [244], authors investigate the effectiveness of linguistically motivated strategies of code mixing in a goal oriented dialogue setting.

Cross-lingual Question Answering systems were extended to dialog systems for railway inquiries [245]. Recently, there has been an attempt to create code-mixed version of goal oriented conversations [246] from the DSTC2 restaurant reservation dataset.

As we have discussed earlier, code-switching is a phenomenon observed in informal scenarios, which implies that it is a suitable setting in conversational speech. In coherence to this thought, there has been work on incorporating and comparing text and speech based features in identifying language at turn level in a dialog [247]. Hence this work focuses on inter-sentential switching as opposed to intra-sentential switching. This work demonstrated the efficacy of i-Vector features in comparison to spectral features for speech segments. While the best performing system is based on text features, this work equips the field to work with speech based features when transcriptions are not available.

4.15. User Interfaces

While languages that are being mixed may share the same script (such as in the case of English and Spanish), this is not true for many language pairs that are frequently code-switched, particularly when the languages are not related to each other. In such cases, users may choose to use the same script to write both languages, or use a mixed script. This has implications not only in how

to process mixed languages, but also on how to display them. [248] presents a study on the interaction between script-mixing and language mixing for Hindi-English and shows that script choice may be used for emphasis, disambiguation and marking whether a word is borrowed or not.

4.16. Optical Character Recognition

[249] extend a standard OCR model to enable transcription of code-switched text by jointly performing transcription with word-level language identification. The model provides significant error reductions in historical texts.

4.17. Improving cross-lingual models

Recently, code-switched text has been used to improve the performance of cross-lingual systems. Code-switching is seen as a bridge to anchor representations in different languages so that they can come closer in a common space and lead to improved performance in cross-lingual NLP tasks [250]. [251] use code-switched text along with an English language identifier to retrieve documents written in Romanized Hindi. [252] use alternating language modeling by artificially generating code-switched text using phrase alignments between parallel sentences to improve performance on cross-lingual tasks such as XNLI [253]. [250] use a similar approach in which they synthesize random code-switched sentences in multiple languages to improve zero-shot performance on XNLI.

5. Evaluation of Code-switched Systems

Much of the progress in a new field can be shaped by shared tasks in which common datasets are released and participants compete to build systems for a specific task. There have been several shared tasks conducted for code-switched text processing, and a few shared tasks for code-switched speech processing over the last few years. Shared tasks for code-switched NLP have included Language Identification [254, 68, 255], transliterated search [256], code-mixed entity extraction [257], mixed script information retrieval [258, 259], POS tagging [260],

36

Named Entity Recognition [75], Sentiment Analysis [229] and Question Answering [90, 91]. There have been fewer shared task for code-switched speech processing, however, the Blizzard challenge 2014 had a code-switched speech synthesis task [261], and code-switched ASR challenges have been conducted for Mandarin English [45, 262]. Recently, a spoken Language Identification challenge was conducted for inter and intra-utterance LID [66] in three code-switched language pairs.

Each of these shared task have spurred research in their respective sub-areas of code-switched speech and NLP. However, it is not clear how well these individual models can generalize across different tasks and language pairs. To address this gap, benchmarks for evaluating code-switching across different NLP have been proposed.

The GLUECoS benchmark [263] consists of 11 datasets spanning different tasks for code-switching across two language pairs Spanish-English and Hindi-English, including a new task for code-switching, Natural Language Inference (NLI). The GLUECoS benchmark aims to add more tasks to evaluate the general language understanding capabilities of models, including tasks such as Question Answering, Natural Language Generation and Summarization, Machine Translation and NLI. The LINCE benchmark [264] consists of 10 datasets across 5 language pairs. The tasks include LID, NER, POS tagging and Sentiment Analysis.

Evaluations conducted on the benchmarks described above indicate that massively multilingual contextual language models such as multilingual BERT [101] outperform cross-lingual models and other task-specific models. These models can be further improved by adding synthetic code-switched data to pre-training, as shown in [263]. While models on some word level tasks such as Language Identification and Named Entity Recognition reach high accuracy, the performance on harder tasks like Sentiment Analysis, Question Answering and NLI is much worse and there is a large gap between the performance of models on monolingual tasks compared to code-switched tasks. This indicates that massive multilingual models do not perform as well on code-switching as

they do on monolingual or even cross-lingual tasks. However, pre-training or fine-tuning such models on synthetic code-switched data in the absence of real code-switched data seems to a be promising future direction.

6. Challenges and Future Directions

Although code-switching is a persistent phenomenon through out the whole world, access to data will always be limited. Monolingual corpora will always be easier to find as monolingual discourse is more common in formal environments and hence more likely to be archived. Code-switching data, by its nature of being used in more informal contexts, is less likely to be archived and hence harder to find as training data. As code-switching is more likely to be used in less task specific contexts, with less explicit function it may also be more difficult to label such data.

Most code-switching studies focus on pairs with one high resource language (e.g. English, Spanish, MSA, Putonghua) and a lower resource language, realistically the position is much more complex than that. Although we consider Hinglish data low resourced, there are many other Northern Indian languages that are code-switched with Hindi and access to that data is even harder. Thus code-switching studies will inherently always be data starved and our models must therefore expect to work with limited data.

Most current work in code-switching looks at one particular language pair. It is not yet the case that architectures for multiple pairs are emerging, except perhaps within the Indian sub-continent where there are similar usage patterns with English and various regional languages. However it is clear that not all code-switching is the same. Relative fluency, social prestige, topical restrictions and grammatical constraints have quite different effects on code-switching practices thus it is hard to consider general code-switching models over multiple language pairs.

We should also take into the account that as code-switching is more typical in less formal occasions, there are some tasks that are more likely to involve

code-switching that others. Thus we are unlikely to encounter programming languages that use code-switching, but we are much more likely to encounter code-switching in sentiment analysis. Likewise analysis of parliamentary transcripts are more likely to be monolingual, while code-switching is much more likely in social media. Of course its not just the forum that affects the distribution, the topic too may be a factor.

These factors of use of code-switch should influence how we consider development of code-switched models. Although it may be possible to build end-to-end systems where large amounts of code-switching data is available, in well-defined task environments, such models will not have the generalizations we need to cover the whole space. For models that can make use of large amounts of unlabeled data for training, generating synthetic code-switched data may be a promising direction. However, most models of code-switched data generation rely on syntactic constraints and do not take into account sociolinguistic factors that affect code-switched language. Building models that are capable of incorporating these factors could lead to more realistic data generation, which could lead to better models for code-switched speech and NLP.

It is not yet clear yet from the NLP point of view if code-switching analysis should be treated primarily as a translation problem, or be treated as a new language itself. It is however likely as with many techniques in low-resource language processing, exploiting resources from nearby languages will have an advantage. It is common (though not always) that one language involved in code-switching has significant resources (e.g. English, Putonghua, Modern Standard Arabic). Thus transfer learning approaches are likely to offer short term advantages. Also given the advancement of language technologies, particularly due to the rise of massively multilingual models, developing techniques that can work over multiple pairs of code-switched languages may lead to faster development and generalization of the field.

Evaluating code-switched speech and NLP is challenging due to the lack of standardized datasets. Although initial attempts at creating benchmarks have been made, a comprehensive evaluation of code-switched systems across

speech and NLP tasks in many typologically different language pairs is required. Such evaluation benchmarks are even more important due to the prevalence of multilingual models that perform zero-shot cross-lingual transfer well, and are also expected to perform well on code-switched languages.

Speech and language technologies for code-switching is not yet a mature field. It is noted that the references to work in this article are for the most part the beginnings of analysis. They are investigating the raw tools that are necessary in order for the development of full systems. While there has been a lot of work on individual Speech and NLP systems for code-switching, there are no end-to-end systems that can interact in code-switched language with multilingual humans. Specifically we are not yet seeing full end-to-end digital assistants for code-switched interaction, or sentiment analysis for code-switched reviews, or grammar and spelling for code-switched text. This is partly due to lack of data for such end-to-end systems, however, a code-switching intelligent agent has to be more than just the sum of parts that can handle code-switching. To build effective systems that can code-switch, we will also have to leverage the work done in sociolinguistics to understand how, when and why to code-switch.

References

References

[1] R. Hickey, The handbook of language contact, John Wiley & Sons, 2012.

[2] P. Auer, A postscript: Code-switching and social identity, Journal of pragmatics 37 (3) (2005) 403–410.

[3] M. Heller, Negotiations of language choice in montreal, Language and social identity (1982) 108–118.

[4] R. Jacobson, Codeswitching Worldwide. II, Vol. 126, Walter de Gruyter, 2011.

[5] A. Camilleri, Language values and identities: Code switching in secondary classrooms in Malta, Linguistics and education 8 (1) (1996) 85–103.

[6] X. Qian, G. Tian, Q. Wang, Codeswitching in the primary efl classroom in china–two case studies, System 37 (4) (2009) 719–730.

[7] E. Rezvani, H. J. Street, A. E. Rasekh, Code-switching in iranian elementary efl classrooms: An exploratory investigation., English language teaching 4 (1) (2011) 18–25.

[8] M. A. Peabody, Methods for pronunciation assessment in computer aided language learning, Ph.D. thesis, Massachusetts Institute of Technology (2011).

[9] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. Lawrence Zitnick, D. Parikh, Vqa: Visual question answering, in: Proceedings of the IEEE international conference on computer vision, 2015, pp. 2425–2433.

[10] P. Agarwal, A. Sharma, J. Grover, M. Sikka, K. Rudra, M. Choudhury, I may talk in english but gaali toh hindi mein hi denge: A study of english-hindi code-switching and swearing pattern on social networks, in: 2017 9th International Conference on Communication Systems and Networks (COMSNETS), IEEE, 2017, pp. 554–557.

[11] L. Todd, L. Todd, Pidgins and creoles, Routledge, 2003.

[12] J. Arends, P. Muysken, N. Smith, Pidgins and creoles: An introduction, Vol. 15, John Benjamins Publishing, 1995.

[13] M. Sebba, Contact languages: Pidgins and creoles, Macmillan International Higher Education, 1997.

[14] C. Myers-Scotton, Contact linguistics: Bilingual Encounters and Grammatical Outcomes, Oxford University Press on Demand, 2002.

[15] J. J. Gumperz, Discourse strategies, Vol. 1, Cambridge University Press, 1982.

[16] C. Myers-Scotton, Duelling languages: Grammatical structure in codeswitching, Oxford University Press, 1997.

[17] K. Bali, J. Sharma, M. Choudhury, 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, 2014, pp. 116–126.

[18] S. Rijhwani, R. Sequiera, M. Choudhury, K. Bali, C. S. Maddila, Estimating code-switching on twitter with a novel generalized word-level language detection technique, in: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Vol. 1, 2017, pp. 1971–1982.

[19] A. K. Joshi, Processing of sentences with intra-sentential code-switching, in: Proceedings of the 9th conference on Computational linguistics-Volume 1, Academia Praha, 1982, pp. 145–150.

[20] S. Poplack, Syntactic structure and social function of code-switching, Vol. 2, Centro de Estudios Puertorriqueños,[City University of New York], 1978.

[21] S. Poplack, Sometimes ill start a sentence in spanish y termino en espanol: toward a typology of code-switching1, Linguistics 18 (7-8) (1980) 581–618.

[22] D. Sankoff, A formal production-based explanation of the facts of code-switching, Bilingualism: language and cognition 1 (1) (1998) 39–50.

[23] A.-M. Di Sciullo, P. Muysken, R. Singh, Government and code-mixing, Journal of linguistics 22 (1) (1986) 1–24.

[24] H. M. Belazi, E. J. Rubin, A. J. Toribio, Code switching and x-bar theory: The functional head constraint, Linguistic inquiry (1994) 221–237.

[25] M. Sebba, A congruence approach to the syntax of codeswitching, International Journal of Bilingualism 2 (1) (1998) 1–19.

[26] P. Gardner-Chloros, M. Edwards, Assumptions behind grammatical approaches to code-switching: When the blueprint is a red herring, Transactions of the Philological Society 102 (1) (2004) 103–129.

[27] G. Bhat, M. Choudhury, K. Bali, Grammatical constraints on intra-sentential code-switching: From theories to working models, arXiv preprint arXiv:1612.04538.

[28] A. Pratapa, G. Bhat, M. Choudhury, S. Sitaram, S. Dandapat, K. Bali, Language modeling for code-mixing: The role of linguistic theory based synthetic data, in: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, 2018.

[29] U. Lanvers, Language Alternation in infant bilinguals: A developmental approach to codeswitching, International Journal of Bilingualism.

[30] A. Backus, Codeswitching and language change: One thing leads to another?, International Journal of Bilingualism 9 (3-4) (2005) 307–340.

[31] V. Soto, N. Cestero, J. Hirschberg, The role of cognate words, pos tags and entrainment in code-switching., in: Interspeech, 2018, pp. 1938–1942.

[32] S. Hartmann, M. Choudhury, K. Bali, An integrated representation of linguistic and social functions of code-switching, in: Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), 2018.

[33] A. Pratapa, M. Choudhury, Quantitative characterization of code switching patterns in complex multi-party conversations: A case study on hindi movie scripts, in: Proceedings of the 14th International Conference on Natural Language Processing (ICON-2017), 2017, pp. 75–84.

[34] B. Gambäck, A. Das, Comparing the Level of Code-Switching in Corpora, in: LREC, 2016.

[35] R. Barnett, E. Codó, E. Eppler, M. Forcadell, P. Gardner-Chloros, R. Van Hout, M. Moyer, M. C. Torras, M. T. Turell, M. Sebba, et al., The LIDES Coding Manual: A Document for Preparing and Analyzing Language Interaction Data Version 1.1, International Journal of Bilingualism 4 (2) (2000) 131–271.

[36] G. A. Guzman, J. Serigos, B. Bullock, A. J. Toribio, Simple tools for exploring variation in code-switching for linguists, in: Proceedings of the Second Workshop on Computational Approaches to Code Switching, 2016, pp. 12–20.

[37] G. Guzmán, J. Ricard, J. Serigos, B. E. Bullock, A. J. Toribio, Metrics for modeling Code-Switching across Corpora, Proc. Interspeech 2017 (2017) 67–71.

[38] G. A. Guzmán, J. Ricard, J. Serigos, B. Bullock, A. J. Toribio, Moving code-switching research toward more empirically grounded methods., in: CDH@ TLT, 2017, pp. 1–9.

[39] B. Bullock, W. Guzmán, J. Serigos, V. Sharath, A. J. Toribio, Predicting the presence of a matrix language in code-switching, in: Proceedings of the third workshop on computational approaches to linguistic code-switching, 2018, pp. 68–75.

[40] B. E. Bullock, G. A. Guzmán, J. Serigos, A. J. Toribio, Should code-switching models be asymmetric?, in: Interspeech, 2018, pp. 2534–2538.

[41] D.-C. Lyu, T.-P. Tan, E. S. Chng, H. Li, Seame: A Mandarin-English Code-Switching Speech Corpus in South-East Asia, in: Eleventh Annual Conference of the International Speech Communication Association, 2010.

[42] G. Lee, T.-N. Ho, E.-S. Chng, H. Li, A review of the mandarin-english code-switching corpus: Seame, in: 2017 International Conference on Asian Language Processing (IALP), IEEE, 2017, pp. 210–213.

[43] Y. Li, Y. Yu, P. Fung, A Mandarin-English Code-Switching Corpus., in: LREC, 2012, pp. 2515–2519.

[44] H.-P. Shen, C.-H. Wu, Y.-T. Yang, C.-S. Hsu, Cecos: A chinese-english code-switching speech database, in: 2011 International Conference on Speech Database and Assessments (Oriental COCOSDA), IEEE, 2011, pp. 120–123.

[45] D. Wang, Z. Tang, D. Tang, Q. Chen, Oc16-ce80: A chinese-english mixlingual database and a speech recognition baseline, in: 2016 Conference of The Oriental Chapter of International Committee for Coordination and Standardization of Speech Databases and Assessment Techniques (O-COCOSDA), IEEE, 2016, pp. 84–88.

[46] J. Y. Chan, P. Ching, T. Lee, Development of a Cantonese-English Code-Mixing Speech Corpus, in: Ninth European Conference on Speech Communication and Technology, 2005.

[47] D.-C. Lyu, R.-Y. Lyu, Y.-c. Chiang, C.-N. Hsu, Speech Recognition on Code-Switching among the Chinese Dialects, in: Acoustics, Speech and Signal Processing, 2006. ICASSP 2006 Proceedings. 2006 IEEE International Conference on, Vol. 1, IEEE, 2006, pp. I–I.

[48] S. Nakayama, T. Kano, Q. T. Do, S. Sakti, S. Nakamura, Japanese-english code-switching speech data construction, in: 2018 Oriental COCOSDA-International Conference on Speech Database and Assessments, IEEE, 2018, pp. 67–71.

[49] M. Deuchar, P. Davies, J. Herring, M. C. P. Couto, D. Carter, Building Bilingual Corpora, Advances in the Study of Bilingualism (2014) 93–111.

[50] J. C. Franco, T. Solorio, Baby-steps towards building a Spanglish Language Model, in: International Conference on Intelligent Text Processing and Computational Linguistics, Springer, 2007, pp. 75–84.

[51] V. Ramanarayanan, D. Suendermann-Oeft, Jee haan, I'd like both, por favor: Elicitation of a Code-Switched Corpus of Hindi–English and Spanish–English Human–Machine Dialog, Proc. Interspeech 2017 (2017) 47–51.

[52] S. Sivasankaran, B. M. L. Srivastava, S. Sitaram, K. Bali, M. Choudhury, Phone Merging for Code-Switched Speech Recognition, in: Proceedings of the Third Workshop on Computational Approaches to Linguistic Code-Switching, 2018, pp. 11–19.

[53] G. Sreeram, K. Dhawan, R. Sinha, Hindi-English Code-Switching Speech Corpus, arXiv preprint arXiv:1810.00662.

[54] A. Pandey, B. M. L. Srivastava, S. V. Gangashetty, Adapting Monolingual Resources for Code-Mixed Hindi-English Speech Recognition, in: Asian Language Processing (IALP), 2017 International Conference on, IEEE, 2017, pp. 218–221.

[55] A. Dey, P. Fung, A Hindi-English Code-Switching Corpus., in: LREC, 2014, pp. 2410–2413.

[56] B. H. Ahmed, T.-P. Tan, Automatic Speech Recognition of Code Switching Speech using 1-best Rescoring, in: Asian Language Processing (IALP), 2012 International Conference on, IEEE, 2012, pp. 137–140.

[57] I. Hamed, M. Elmahdy, 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), 2018.

[58] D. MOHDEB-AMAZOUZ, A.-D. Martine, L. LAMEL, Arabic-French Code-Switching across Maghreb Arabic dialects: A Quantitative Analysis.

[59] D. Amazouz, M. Adda-Decker, L. Lamel, The French-Algerian Code-Switching Triggered Audio Corpus (FACST), in: LREC 2018 11th edition of the Language Resources and Evaluation Conference,, 2018.

[60] Ö. Çetinoğlu, A code-switching corpus of turkish-german conversations, in: Proceedings of the 11th Linguistic Annotation Workshop, 2017, pp. 34–40.

[61] E. Yilmaz, J. Dijkstra, H. Velde, H. Heuvel, D. van Leeuwen, Longitudinal Speaker Clustering and Verification Corpus with Code-Switching Frisian-Dutch Speech.

[62] E. van der Westhuizen, T. Niesler, Automatic Speech Recognition of English-isizulu Code-Switched Speech from South African Soap Operas, Procedia Computer Science 81 (2016) 121–127.

[63] E. van der Westhuizen, T. Niesler, A first south african corpus of multilingual code-switched soap opera speech, in: Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), 2018.

[64] E. Yılmaz, A. Biswas, E. van der Westhuizen, F. de Wet, T. Niesler, Building a Unified Code-Switching ASR System for South African Languages, arXiv preprint arXiv:1807.10949.

[65] T. I. Modipa, M. H. Davel, F. De Wet, Implications of Sepedi/English Code Switching for ASR Systems.

[66] S. Shah, S. Sitaram, R. Mehta, First workshop on speech processing for code-switching in multilingual communities: Shared task on code-switched spoken language identification.

[67] A. Baby, A. L. Thomas, H. Myrthy, Resources for Indian Languages, in: Proceedings of Text, Speech and Dialogue, 2016.

[68] T. Solorio, E. Blair, S. Maharjan, S. Bethard, M. Diab, M. Ghoneim, A. Hawwari, F. AlGhamdi, J. Hirschberg, A. Chang, et al., 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, 2014, pp. 62–72.

[69] R. Sequiera, M. Choudhury, P. Gupta, P. Rosso, S. Kumar, S. Banerjee, S. K. Naskar, S. Bandyopadhyay, G. Chittaranjan, A. Das, et al., Overview of FIRE-2015 Shared Task on Mixed Script Information Retrieval.

[70] U. Barman, A. Das, J. Wagner, 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, 2014, pp. 13–23.

[71] A. Das, B. Gambäck, Identifying Languages at the Word level in Code-Mixed Indian Social Media Text.

[72] J. Patro, B. Samanta, S. Singh, A. Basu, P. Mukherjee, M. Choudhury, A. Mukherjee, All that is English may be Hindi: Enhancing Language Identification through Automatic Ranking of the Likeliness of Word Borrowing in Social Media, in: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, 2017, pp. 2264–2274.

[73] D. Jurgens, Y. Tsvetkov, D. Jurafsky, Incorporating Dialectal Variability for Socially Equitable Language Identification, in: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), Vol. 2, 2017, pp. 51–57.

[74] S. Maharjan, E. Blair, S. Bethard, T. Solorio, Developing language-tagged corpora for code-switching tweets, in: Proceedings of The 9th Linguistic Annotation Workshop, 2015, pp. 72–84.

[75] G. Aguilar, F. AlGhamdi, V. Soto, M. Diab, J. Hirschberg, 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, 2018, pp. 138–147.

[76] V. Singh, D. Vijay, S. S. Akhtar, M. Shrivastava, Named Entity Recognition for Hindi-English Code-Mixed Social Media Text, in: Proceedings of the Seventh Named Entities Workshop, 2018, pp. 27–35.

[77] Y. Vyas, S. Gella, J. Sharma, K. Bali, M. Choudhury, POS Tagging of English-Hindi Code-Mixed Social Media Content, in: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2014, pp. 974–979.

[78] S. Ghosh, S. Ghosh, D. Das, Part-of-speech tagging of code-mixed social media text, in: Proceedings of the Second Workshop on Computational Approaches to Code Switching, 2016, pp. 90–97.

[79] Ö. Çetinoğlu, Ç. Çöltekin, Part of Speech Annotation of a Turkish-German Code-Switching Corpus, in: Proceedings of the 10th Linguistic Annotation Workshop held in conjunction with ACL 2016 (LAW-X 2016), 2016, pp. 120–130.

[80] T. Solorio, Y. Liu, Part-Of-Speech Tagging for English-Spanish Code-Switched Text, in: Proceedings of the Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, 2008, pp. 1051–1060.

[81] A. Jamatia, B. Gambäck, 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, 2015, pp. 239–248.

[82] A. Jamatia, B. Gambäck, A. Das, Collecting and annotating indian social media code-mixed corpora, in: International Conference on Intelligent Text Processing and Computational Linguistics, Springer, 2016, pp. 406–417.

[83] V. Soto, J. Hirschberg, Crowdsourcing universal part-of-speech tags for code-switching.

[84] M. Diab, M. Ghoneim, A. Hawwari, F. AlGhamdi, N. AlMarwani, M. Al-Badrashiny, Creating a large multi-layered representational repository of linguistic code switched arabic data, LREC 2016.

[85] N. Partanen, K. Lim, M. Rießler, T. Poibeau, Dependency parsing of code-switching data with cross-lingual feature representations, in: Proceedings of the Fourth International Workshop on Computatinal Linguistics of Uralic Languages, 2018, pp. 1–17.

[86] I. Bhat, R. A. Bhat, M. Shrivastava, D. Sharma, Joining hands: Exploiting monolingual treebanks for parsing of code-mixing data, in: Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers, Vol. 2, 2017, pp. 324–330.

[87] A. Sharma, S. Gupta, R. Motlani, P. Bansal, M. Shrivastava, R. Mamidi, D. M. Sharma, Shallow parsing pipeline for hindi-english code-mixed social media text, in: Proceedings of NAACL-HLT, 2016, pp. 1340–1345.

[88] L. Duong, H. Afshar, D. Estival, G. Pink, P. Cohen, M. Johnson, Multilingual semantic parsing and code-switching, in: Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017), 2017, pp. 379–389.

[89] K. C. Raghavi, M. Chinnakotla, M. Shrivastava, "Answer ka type kya he?" Learning to Classify Questions in Code-Mixed Language.

[90] K. Chandu, E. Loginova, V. Gupta, J. van Genabith, G. Neuman, M. Chinnakotla, E. Nyberg, 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, 2018, pp. 29–38.

[91] K. Chandu, E. Loginova, V. Gupta, J. v. Genabith, G. Neumann, M. Chinnakotla, E. Nyberg, A. W. Black, Code-mixed question answering challenge: Crowd-sourcing data and techniques, in: Third Workshop on Computational Approaches to Linguistic Code-Switching, Association for Computational Linguistics (ACL), 2019, pp. 29–38.

[92] V. Gupta, M. Chinnakotla, M. Shrivastava, Transliteration Better than Translation? Answering Code-mixed Questions over a Knowledge Base, in: Proceedings of the Third Workshop on Computational Approaches to Linguistic Code-Switching, 2018, pp. 39–50.

[93] S. Banerjee, S. K. Naskar, P. Rosso, S. Bandyopadhyay, The First Cross-Script Code-Mixed Question Answering Corpus.

[94] S. Khanuja, S. Dandapat, S. Sitaram, M. Choudhury, A new dataset for natural language inference from code-mixed conversations, in: Proceedings of the The 4th Workshop on Computational Approaches to Code Switching, 2020, pp. 9–16.

[95] K. Chakma, A. Das, Cmir: A corpus for evaluation of code mixed information retrieval of hindi-english tweets, Computación y Sistemas 20 (3) (2016) 425–434.

[96] A. Bohra, D. Vijay, V. Singh, S. S. Akhtar, M. Shrivastava, A dataset of hindi-english code-mixed social media text for hate speech detection, in: Proceedings of the second workshop on computational modeling of peoples opinions, personality, and emotions in social media, 2018, pp. 36–41.

[97] D. Vijay, A. Bohra, V. Singh, S. S. Akhtar, M. Shrivastava, A dataset for detecting irony in hindi-english code-mixed social media text., in: EM-SASW@ ESWC, 2018, pp. 38–46.

[98] S. Thara, P. Poornachandran, Code-mixing: A brief survey, in: 2018 International Conference on Advances in Computing, Communications and Informatics (ICACCI), IEEE, 2018, pp. 2382–2388.

[99] Ö. Çetinoğlu, S. Schulz, N. T. Vu, Challenges of computational processing of code-switching, arXiv preprint arXiv:1610.02213.

[100] A. Pratapa, G. Bhat, M. Choudhury, S. Sitaram, S. Dandapat, K. Bali, Language modeling for code-mixing: The role of linguistic theory based synthetic data, in: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, 2018.

[101] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Multilingual bert (2018).

[102] J. Y. Chan, P. Ching, T. Lee, H. M. Meng, Detection of language boundary in code-switching utterances by bi-phone probabilities, in: International Symposium on Chinese Spoken Language Processing, IEEE, 2004.

[103] J. Weiner, N. T. Vu, D. Telaar, F. Metze, T. Schultz, D.-C. Lyu, E.-S. Chng, H. Li, Integration of Language Identification into a recognition system for spoken conversations containing code-switches, in: Spoken Language Technologies for Under-Resourced Languages, 2012.

[104] J. Y. Chan, H. Cao, P. Ching, T. Lee, Automatic recognition of Cantonese-English code-mixing speech, International Journal of Computational Linguistics & Chinese Language Processing, Volume 14, Number 3, September 2009.

[105] S. Yu, S. Hu, S. Zhang, B. Xu, Chinese-English bilingual speech recognition, in: Proceedings of International Conference on Natural Language Processing and Knowledge Engineering, IEEE, 2003.

[106] N. T. Vu, D.-C. Lyu, J. Weiner, D. Telaar, T. Schlippe, F. Blaicher, E.-S. Chng, T. Schultz, H. Li, A first speech recognition system for Mandarin-English code-switch conversational speech, in: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2012.

[107] T. Lyudovyk, V. Pylypenko, Code-switching Speech Recognition for closely related languages, in: Spoken Language Technologies for Under-Resourced Languages, 2014.

[108] C.-F. Yeh, L.-S. Lee, An improved framework for recognizing highly imbalanced bilingual code-switched lectures with Cross-language Acoustic modeling and frame-level language identification, IEEE Transactions on Audio, Speech, and Language Processing(ICASSP) 2015.

[109] C. M. White, S. Khudanpur, J. K. Baker, An investigation of acoustic models for multilingual code-switching, in: Ninth Annual Conference of the International Speech Communication Association, 2008.

[110] K. Bhuvanagiri, S. Kopparapu, An approach to mixed language Automatic Speech Recognition, Oriental COCOSDA, Kathmandu, Nepal, 2010.

[111] K. Bhuvanagirir, S. K. Kopparapu, Mixed language speech recognition without explicit identification of language, American Journal of Signal Processing.

[112] K. Taneja, S. Guha, P. Jyothi, B. Abraham, Exploiting monolingual speech corpora for code-mixed speech recognition, Proc. Interspeech 2019 (2019) 2150–2154.

[113] Y. Long, Y. Li, Q. Zhang, S. Wei, H. Ye, J. Yang, Acoustic data augmentation for mandarin-english code-switching speech recognition, Applied Acoustics 161 (2020) 107175.

[114] E. Yılmaz, M. McLaren, H. van den Heuvel, D. A. van Leeuwen, Semi-supervised acoustic model training for speech with code-switching, Speech Communication 105 (2018) 12–22.

[115] E. Yılmaz, S. Cohen, X. Yue, D. van Leeuwen, H. Li, Multi-graph decoding for code-switching asr, arXiv preprint arXiv:1906.07523.

[116] G. I. Winata, S. Cahyawijaya, Z. Lin, Z. Liu, P. Xu, P. Fung, Meta-transfer learning for code-switched speech recognition, arXiv preprint arXiv:2004.14228.

[117] H. Zhang, H. Xu, V. T. Pham, H. Huang, E. S. Chng, Monolingual data selection analysis for english-mandarin hybrid code-switching speech recognition, arXiv preprint arXiv:2006.07094.

[118] E. Yılmaz, H. van den Heuvel, D. van Leeuwen, Investigating bilingual deep neural networks for automatic recognition of code-switching frisian speech, Procedia Computer Science 81 (2016) 159–166.

[119] E. Yilmaz, H. Heuvel, D. A. van Leeuwen, Exploiting untranscribed broadcast data for improved code-switching detection.

[120] E. Yılmaz, H. v. d. Heuvel, D. A. van Leeuwen, Acoustic and textual data augmentation for improved asr of code-switching speech, arXiv preprint arXiv:1807.10945.

[121] A. Biswas, E. Yılmaz, F. de Wet, E. van der Westhuizen, T. Niesler, Semi-supervised development of asr systems for multilingual code-switched speech in under-resourced languages, arXiv preprint arXiv:2003.03135.

[122] P. Guo, H. Xu, L. Xie, E. S. Chng, Study of semi-supervised approaches to improving english-mandarin code-switching speech recognition, arXiv preprint arXiv:1806.06200.

[123] S. Nakayama, A. Tjandra, S. Sakti, S. Nakamura, Speech chain for semi-supervised learning of japanese-english code-switching asr and tts, in: 2018 IEEE Spoken Language Technology Workshop (SLT), IEEE, 2018, pp. 182–189.

[124] S. Nakayama, A. Tjandra, S. Sakti, S. Nakamura, Zero-shot code-switching asr and tts with multilingual machine speech chain, in: 2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), IEEE, 2019, pp. 964–971.

[125] X. Song, Y. Liu, D. Yang, Y. Zou, A multi-task learning approach for mandarin-english code-switching conversational speech recognition, in: International Symposium on Intelligence Computation and Applications, Springer, 2017, pp. 102–111.

[126] X. Song, Y. Zou, S. Huang, S. Chen, Y. Liu, Investigating multi-task learning for automatic speech recognition with code-switching between mandarin and english, in: 2017 International Conference on Asian Language Processing (IALP), IEEE, 2017, pp. 27–30.

[127] G. I. Winata, A. Madotto, C.-S. Wu, P. Fung, Towards end-to-end automatic code-switching speech recognition, arXiv preprint arXiv:1810.12620.

[128] C. Shan, C. Weng, G. Wang, D. Su, M. Luo, D. Yu, L. Xie, Investigating end-to-end speech recognition for mandarin-english code-switching.

[129] S. Nakayama, T. Kano, A. Tjandra, S. Sakti, S. Nakamura, Recognition and translation of code-switching speech utterances.

[130] Y. Khassanov, H. Xu, V. T. Pham, Z. Zeng, E. S. Chng, C. Ni, B. Ma, Constrained output embeddings for end-to-end code-switching speech recognition with only monolingual data, arXiv preprint arXiv:1904.03802.

[131] K. Li, J. Li, G. Ye, R. Zhao, Y. Gong, Towards code-switching asr for end-to-end ctc models, in: ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2019, pp. 6076–6080.

[132] S. Zhang, J. Yi, Z. Tian, J. Tao, Y. Bai, Rnn-transducer with language bias for end-to-end mandarin-english code-switching speech recognition, arXiv preprint arXiv:2002.08126.

[133] N. Luo, D. Jiang, S. Zhao, C. Gong, W. Zou, X. Li, Towards end-to-end code-switching speech recognition, arXiv preprint arXiv:1810.13091.

[134] Z. Zeng, Y. Khassanov, V. T. Pham, H. Xu, E. S. Chng, H. Li, On the end-to-end solution to mandarin-english code-switching speech recognition, arXiv preprint arXiv:1811.00241.

[135] K. Li, J. Li, G. Ye, R. Zhao, Y. Gong, Towards code-switching asr for end-to-end ctc models, in: ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2019, pp. 6076–6080.

[136] S. Shah, B. Abraham, S. Sitaram, V. Joshi, et al., Learning to recognize code-switched speech without forgetting monolingual speech recognition, arXiv preprint arXiv:2006.00782.

[137] G. Reddy Madhumani, S. Shah, B. Abraham, V. Joshi, S. Sitaram, Learning not to discriminate: Task agnostic learning for improving monolingual and code-switched speech recognition, arXiv (2020) arXiv–2006.

[138] B. M. L. Srivastava, S. Sitaram, Homophone Identification and Merging for Code-switched Speech Recognition, Proeedings of Interspeech 2018.

[139] F. Weng, H. Bratt, L. Neumeyer, A. Stolcke, A study of multilingual speech recognition, in: Fifth European Conference on Speech Communication and Technology, 1997.

[140] Y. Li, P. Fung, Code-switch language model with inversion constraints for mixed language speech recognition, Proceedings of COLING 2012 (2012) 1671–1680.

[141] Y. Li, P. Fung, Code switch language modeling with functional head constraint, in: 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2014, pp. 4913–4917.

[142] Y. Li, P. Fung, Language modeling for mixed language speech recognition using weighted phrase extraction., in: Interspeech, 2013, pp. 2599–2603.

[143] A. Baheti, S. Sitaram, M. Choudhury, K. Bali, Curriculum design for code-switching: Experiments with language identification and language modeling with deep neural networks, Proceedings of International Conference on Natural Language Processing (ICON), 2017.

[144] J. Gebhardt, Speech recognition on english-mandarin code-switching data using factored language models, Ph.D. thesis, MS thesis, Department of Informatics, Karlsruhe Institute of Technology (2011).

[145] H. Adel, N. T. Vu, F. Kraus, T. Schlippe, H. Li, T. Schultz, Recurrent neural network language modeling for code switching conversational speech, in: 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, IEEE, 2013, pp. 8411–8415.

[146] H. Adel, N. T. Vu, T. Schultz, Combination of recurrent neural networks and factored language models for code-switching language modeling, in: Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), Vol. 2, 2013, pp. 206–211.

[147] H. Adel, K. Kirchhoff, D. Telaar, N. T. Vu, T. Schlippe, T. Schultz, Features for factored language models for code-switching speech, in: Spoken Language Technologies for Under-Resourced Languages, 2014.

[148] N. T. Vu, T. Schultz, Exploration of the impact of maximum entropy in recurrent neural network language models for code-switching speech, in: Proceedings of The First Workshop on Computational Approaches to Code Switching, 2014, pp. 34–41.

[149] H. Adel, N. T. Vu, K. Kirchhoff, D. Telaar, T. Schultz, Syntactic and semantic features for code-switching factored language models, IEEE/ACM transactions on audio, speech, and language Processing 23 (3) (2015) 431–440.

[150] N. T. Vu, H. Adel, T. Schultz, An investigation of code-switching attitude dependent language modeling, in: International Conference on Statistical Language and Speech Processing, Springer, 2013, pp. 297–308.

[151] E. van der Westhuizen, T. Niesler, Synthesising isizulu-english code-switch bigrams using word embeddings., in: INTERSPEECH, 2017, pp. 72–76.

[152] S. Garg, T. Parekh, P. Jyothi, Code-switched language models using dual rnns and same-source pretraining, arXiv preprint arXiv:1809.01962.

[153] F. Blaicher, Smt-based text generation for code-switching language models, Ph.D. thesis, Masters thesis, Cognitive Systems Lab (CSL), Karlsruhe Insitutite of (2011).

[154] G. I. Winata, A. Madotto, C.-S. Wu, P. Fung, Code-switched language models using neural based synthetic data from parallel sentences, arXiv preprint arXiv:1909.08582.

[155] C.-T. Chang, S.-P. Chuang, H.-Y. Lee, Code-switching sentence generation by generative adversarial networks and its application to data augmentation, arXiv preprint arXiv:1811.02356.

[156] K. Chandu, T. Manzini, S. Singh, A. W. Black, Language informed modeling of code-switched text, in: Proceedings of the Third Workshop on Computational Approaches to Linguistic Code-Switching, 2018, pp. 92–97.

[157] H. Gonen, Y. Goldberg, Language modeling for code-switching: Evaluation, integration of monolingual data, and discriminative training, arXiv preprint arXiv:1810.11895.

[158] Z. Zeng, H. Xu, T. Y. Chong, E.-S. Chng, H. Li, Improving n-gram language modeling for code-switching speech recognition, in: 2017 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), IEEE, 2017, pp. 1596–1601.

[159] G. I. Winata, A. Madotto, C.-S. Wu, P. Fung, Code-switching language modeling using syntax-aware multi-task learning, arXiv preprint arXiv:1805.12070.

[160] G. Lee, H. Li, Modeling code-switch languages using bilingual parallel corpus, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020, pp. 860–870.

[161] P. E. Piccinini, M. Garellek, Prosodic cues to monolingual versus code-switching sentences in english and spanish, in: Proceedings of the 7th Speech Prosody Conference, 2014, pp. 885–889.

[162] S. Rallabandi, A. W. Black, On building mixed lingual speech synthesis systems, Proceedings of Interspeech.

[163] E. Yılmaz, H. van den Heuvel, D. van Leeuwen, Code-switching detection using multilingual dnns, in: 2016 IEEE Spoken Language Technology Workshop (SLT), IEEE, 2016, pp. 610–616.

[164] E. Yılmaz, H. v. d. Heuvel, D. A. van Leeuwen, Code-switching detection with data-augmented acoustic and language models, arXiv preprint arXiv:1808.00521.

[165] D.-C. Lyu, R.-Y. Lyu, C.-L. Zhu, M.-T. Ko, Language identification in code-switching speech using word-based lexical model, in: 2010 7th International Symposium on Chinese Spoken Language Processing, IEEE, 2010, pp. 460–464.

[166] D.-C. Lyu, T.-P. Tan, E.-S. Chng, H. Li, An analysis of a mandarin-english code-switching speech corpus: Seame, Age 21 (2010) 25–8.

[167] K. R. Mabokela, M. J. Manamela, M. Manaileng, Modeling code-switching speech on under-resourced languages for language identification, in: Spoken Language Technologies for Under-Resourced Languages, 2014.

[168] L. M. Tomokiyo, A. W. Black, K. A. Lenzo, Foreign accents in synthetic speech: Development and Evaluation, in: Ninth European Conference on Speech Communication and Technology, 2005.

[169] N. Campbell, Talking foreign - Concatenative Speech Synthesis and the Language Barrier, in: Seventh European Conference on Speech Communication and Technology, 2001.

[170] L. Badino, C. Barolo, S. Quazza, Language independent phoneme mapping for foreign TTS, in: Fifth ISCA Workshop on Speech Synthesis, 2004.

[171] M. Mashimo, T. Toda, K. Shikano, N. Campbell, Evaluation of cross-language voice conversion based on GMM and STRAIGHT.

[172] J. Latorre, K. Iwano, S. Furui, Polyglot synthesis using a mixture of Monolingual corpora, in: International Conference on Acoustics, Speech, and Signal Processing, 2005. Proceedings(ICASSP), IEEE, 2005.

[173] S. Sitaram, S. K. Rallabandi, S. Black, Experiments with cross-lingual systems for synthesis of Code-Mixed text, in: 9th ISCA Speech Synthesis Workshop, 2017.

[174] N. K. Elluru, A. Vadapalli, R. Elluru, H. Murthy, K. Prahallad, Is word-to-phone mapping better than phone-phone mapping for handling English words?, in: Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, 2013.

[175] S. K. Rallabandi, A. Vadapalli, S. Achanta, S. Gangashetty, IIIT Hyderabad's submission to the Blizzard Challenge 2015, in: Proceedings of Blizzard Challenge 2015, ISCA, 2015.

[176] A. L. Thomas, A. Prakash, A. Baby, H. A. Murthy, Code-switching in indic speech synthesisers., in: Interspeech, 2018, pp. 1948–1952.

[177] Y. Cao, X. Wu, S. Liu, J. Yu, X. Li, Z. Wu, X. Liu, H. Meng, End-to-end code-switched tts with mix of monolingual recordings, in: ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2019, pp. 6935–6939.

[178] Y. Cao, S. Liu, X. Wu, S. Kang, P. Liu, Z. Wu, X. Liu, D. Su, D. Yu, H. Meng, Code-switched speech synthesis using bilingual phonetic posteriorgram with only monolingual corpora, in: ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2020, pp. 7619–7623.

[179] X. Zhou, E. Yılmaz, Y. Long, Y. Li, H. Li, Multi-encoder-decoder transformer for code-switching speech recognition, arXiv preprint arXiv:2006.10414.

[180] C. Lignos, M. Marcus, Toward web-scale analysis of codeswitching, in: 87th Annual Meeting of the Linguistic Society of America, 2013.

[181] G. Chittaranjan, Y. Vyas, K. Bali, M. Choudhury, Word level Language Identification using CRF: Code-Switching shared task report of MSR India system, in: Proceedings of The First Workshop on Computational Approaches to Code Switching, 2014, pp. 73–79.

[182] M. X. Xia, Codeswitching language identification using subword information enriched word vectors, in: Proceedings of The Second Workshop on Computational Approaches to Code Switching, 2016, pp. 132–136.

[183] D. Nguyen, L. Cornips, Automatic detection of intra-word code-switching, in: Proceedings of the 14th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology, 2016, pp. 82–86.

[184] A. Chanda, D. Das, C. Mazumdar, Unraveling the english-bengali code-mixing phenomenon, in: Proceedings of the Second Workshop on Computational Approaches to Code Switching, 2016, pp. 80–89.

[185] Y.-L. Yeong, T.-P. Tan, Language identification of code switching malay-english words using syllable structure information, in: Spoken Languages Technologies for Under-Resourced Languages, 2010.

[186] H. Elfardy, M. Diab, Token level identification of linguistic code switching, in: Proceedings of COLING 2012: Posters, 2012, pp. 287–296.

[187] Y.-L. Yeong, T.-P. Tan, Applying grapheme, word, and syllable information for language identification in code switching sentences, in: 2011 International Conference on Asian Language Processing, IEEE, 2011, pp. 111–114.

[188] Y.-L. Yeong, T.-P. Tan, Language identification of code switching sentences and multilingual sentences of under-resourced languages by using multi structural word information, in: Fifteenth Annual Conference of the International Speech Communication Association, 2014.

[189] R. Shirvani, M. Piergallini, G. S. Gautam, M. Chouikha, The howard university system submission for the shared task in language identification in spanish-english codeswitching, in: Proceedings of the second workshop on computational approaches to code switching, 2016, pp. 116–120.

[190] N. Dongen, Analysis and prediction of dutch-english code-switching in dutch social media messages, Master's thesis, Universiteit van Amsterdam, Amsterdam, Netherlands.

[191] P. Shrestha, Codeswitching detection via lexical features in conditional random fields, in: Proceedings of the Second Workshop on Computational Approaches to Code Switching, 2016, pp. 121–126.

[192] Y. Samih, W. Maier, Detecting code-switching in moroccan arabic social media, SocialNLP@ IJCAI-2016, New York.

[193] Y. Samih, S. Maharjan, M. Attia, L. Kallmeyer, T. Solorio, Multilingual code-switching identification via lstm recurrent neural networks, in: Proceedings of the Second Workshop on Computational Approaches to Code Switching, 2016, pp. 50–59.

[194] U. K. Sikdar, B. Gambäck, Language identification in code-switched text using conditional random fields and babelnet, in: Proceedings of the Second Workshop on Computational Approaches to Code Switching, 2016, pp. 127–131.

[195] A. Jaech, G. Mulcaire, M. Ostendorf, N. A. Smith, A neural model for language identification in code-switched tweets, in: Proceedings of The Second Workshop on Computational Approaches to Code Switching, 2016, pp. 60–64.

[196] J. C. Chang, C.-C. Lin, Recurrent-neural-network for language detection on twitter code-switching corpus, arXiv preprint arXiv:1412.4314.

[197] N. Jain, R. A. Bhat, Language identification in code-switching scenario, in: Proceedings of the First Workshop on Computational Approaches to Code Switching, 2014, pp. 87–93.

[198] A. Chanda, D. Das, C. Mazumdar, Columbia-jadavpur submission for emnlp 2016 code-switching workshop shared task: System description, in:

Proceedings of the Second Workshop on Computational Approaches to Code Switching, 2016, pp. 112–115.

[199] H. Jhamtani, S. K. Bhogi, V. Raychoudhury, Word-level Language Identification in Bi-lingual code-switched texts, in: Proceedings of the 28th Pacific Asia Conference on Language, Information and Computing, 2014.

[200] B. King, S. Abney, Labeling the languages of words in Mixed-Language documents using weakly supervised methods, in: Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2013.

[201] M. Z. Ansari, S. Khan, T. Amani, A. Hamid, S. Rizvi, Analysis of part of speech tags in language identification of code-mixed text, in: Advances in Computing and Intelligent Systems, Springer, 2020, pp. 417–425.

[202] M. Attia, Y. Samih, W. Maier, Ghht at calcs 2018: Named Entity Recognition for Dialectal Arabic using Neural Networks, in: Proceedings of the Third Workshop on Computational Approaches to Linguistic Code-Switching, 2018.

[203] P. Geetha, K. Chandu, A. W. Black, Tackling Code-Switched NER: Participation of CMU, in: Proceedings of the Third Workshop on Computational Approaches to Linguistic Code-Switching, 2018.

[204] G. Aguilar, S. Maharjan, A. P. L. Monroy, T. Solorio, A Multi-task approach for Named Entity Recognition in Social Media data, in: Proceedings of the 3rd Workshop on Noisy User-generated Text, 2017.

[205] A. Zirikly, M. Diab, Named Entity Recognition for Arabic Social Media, in: Proceedings of the 1st Workshop on Vector Space Modeling for Natural Language Processing, 2015.

[206] J. L. C. Zea, J. E. O. Luna, C. Thorne, G. Glavaš, Spanish NER with Word Representations and Conditional Random Fields, in: Proceedings of the Sixth Named Entity Workshop, 2016.

[207] D. Etter, F. Ferraro, R. Cotterell, O. Buzek, B. Van Durme, Nerit: Named Entity Recognition for Informal Text.

[208] C. Sabty, A. Sherif, M. Elmahdy, S. Abdennadher, Techniques for named entity recognition on arabic-english code-mixed data.

[209] A. Akbik, D. Blythe, R. Vollgraf, Contextual string embeddings for sequence labeling, in: COLING 2018, 27th International Conference on Computational Linguistics, 2018, pp. 1638–1649.

[210] G. I. Winata, C.-S. Wu, A. Madotto, P. Fung, Bilingual character representation for efficiently addressing out-of-vocabulary words in code-switching named entity recognition, arXiv preprint arXiv:1805.12061.

[211] C. Wang, K. Cho, D. Kiela, Code-switched named entity recognition with embedding attention, in: Proceedings of the Third Workshop on Computational Approaches to Linguistic Code-Switching, 2018, pp. 154–158.

[212] G. I. Winata, Z. Lin, P. Fung, Learning multilingual meta-embeddings for code-switching named entity recognition, in: Proceedings of the 4th Workshop on Representation Learning for NLP (RepL4NLP-2019), 2019, pp. 181–186.

[213] G. I. Winata, Z. Lin, J. Shin, Z. Liu, P. Fung, Hierarchical meta-embeddings for code-switching named entity recognition, arXiv preprint arXiv:1909.08504.

[214] U. Barman, J. Wagner, J. Foster, Part-of-speech tagging of code-mixed social media content: Pipeline, stacking and joint modelling, in: Proceedings of the Second Workshop on Computational Approaches to Code Switching, 2016, pp. 30–39.

[215] R. van der Goot, Ö. Çetinoğlu, Lexical normalization for code-switched data and its effect on pos-tagging, arXiv preprint arXiv:2006.01175.

[216] P. Goyal, M. R. Mital, A. Mukerjee, A Bilingual Parser for Hindi, English and code-switching structures, in: 10th Conference of The European Chapter, 2003.

[217] D. Sharma, K. Vikram, M. R. Mital, A. Mukerjee, A. M. Raina, Saarthaka- An Integrated Discourse Semantic Model for Bilingual Corpora, in: Proceedings of International Conference on Universal Knowledge and Language, 2002.

[218] I. A. Bhat, R. A. Bhat, M. Shrivastava, D. M. Sharma, Universal dependency parsing for hindi-english code-switching, arXiv preprint arXiv:1804.05868.

[219] X. Li, D. Roth, Learning Question Classifiers, in: Proceedings of the 19th international conference on Computational linguistics-Volume 1, Association for Computational Linguistics, 2002.

[220] K. R. Chandu, M. Chinnakotla, A. W. Black, M. Shrivastava, Webshodh: A Code Mixed Factoid Question Answering System for Web, in: International Conference of the Cross-Language Evaluation Forum for European Languages, Springer, 2017.

[221] S. Sekine, R. Grishman, Hindi-English cross-lingual Question-Answering System, ACM Transactions on Asian Language Information Processing (TALIP) 2003.

[222] P. Pingali, J. Jagarlamudi, V. Varma, A Dictionary based approach with Query Expansion to Cross Language Query based Multi-Document Summarization: Experiments in Telugu-English, Citeseer 2008.

[223] G. Neumann, B. Sacaleanu, A Cross Language Question/Answering System for German and English.

[224] B. Magnini, S. Romagnoli, A. Vallin, J. Herrera, A. Peñas, V. Peinado, F. Verdejo, M. de Rijke, The multiple Language Question Answering Track, 2003.

[225] D. Vilares, M. A. Alonso, 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'16), 2016, pp. 4149–4153.

[226] D. Vilares, M. A. Alonso, 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, 2015, pp. 2–8.

[227] D. Vilares, M. A. Alonso, C. Gómez-Rodríguez, Supervised sentiment analysis in multilingual environments, Information Processing & Management 53 (3) (2017) 595–607.

[228] S. Padmaja, S. Fatima, S. Bandu, M. Nikitha, K. Prathyusha, Sentiment extraction from bilingual code mixed social media text, in: Data Engineering and Communication Technology, Springer, 2020, pp. 707–714.

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

[230] S. Mandal, D. Das, Analyzing roles of classifiers and code-mixed factors for sentiment identification, arXiv preprint arXiv:1801.02581.

[231] Y. K. Lal, V. Kumar, M. Dhar, M. Shrivastava, P. Koehn, De-mixing sentiment from code-mixed text, in: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop, 2019, pp. 371–377.

[232] J. Utsav, D. Kabaria, R. Vajpeyi, M. Mina, V. Srivastava, Stance detection in hindi-english code-mixed data, in: Proceedings of the 7th ACM IKDD CoDS and 25th COMAD, 2020, pp. 359–360.

[233] S. Lee, Z. Wang, Emotion in code-switching texts: Corpus construction and analysis, in: Proceedings of the eighth SIGHAN workshop on Chinese language processing, 2015, pp. 91–99.

[234] Z. Wang, S. Lee, S. Li, G. Zhou, Emotion detection in code-switching texts via bilingual and sentimental information, in: Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers), 2015, pp. 763–768.

[235] Z. Wang, S. Y. M. Lee, S. Li, G. Zhou, Emotion analysis in code-switching text with joint factor graph model, IEEE/ACM Transactions on Audio, Speech, and Language Processing 25 (3) (2016) 469–480.

[236] C. Supraja, V. M. Rao, Emotion detection in code-switching text, Emotion.

[237] K. Rajput, R. Kapoor, P. Mathur, P. Kumaraguru, R. R. Shah, et al., Transfer learning for detecting hateful sentiments in code switched language, in: Deep Learning-Based Approaches for Sentiment Analysis, Springer, 2020, pp. 159–192.

[238] A. Khandelwal, N. Kumar, A unified system for aggression identification in english code-mixed and uni-lingual texts, in: Proceedings of the 7th ACM IKDD CoDS and 25th COMAD, 2020, pp. 55–64.

[239] T. Santosh, K. Aravind, Hate speech detection in hindi-english code-mixed social media text, in: Proceedings of the ACM India Joint International Conference on Data Science and Management of Data, 2019, pp. 310–313.

[240] R. M. K. Sinha, A. Thakur, Machine Translation of Bilingual Hindi-English(Hinglish) Text, 10th Machine Translation summit (MT Summit X), Phuket, Thailand.

[241] M. Johnson, M. Schuster, Q. V. Le, M. Krikun, Y. Wu, Z. Chen, N. Thorat, F. Viégas, M. Wattenberg, G. Corrado, et al., Googles multilingual

neural machine translation system: Enabling zero-shot translation, Transactions of the Association for Computational Linguistics 5 (2017) 339–351.

[242] P. Rao, M. Pandya, K. Sabu, K. Kumar, N. Bondale, A study of lexical and prosodic cues to segmentation in a hindi-english code-switched discourse, Proc. Interspeech 2018 (2018) 1918–1922.

[243] A. Bawa, M. Choudhury, K. Bali, Accommodation of conversational code-choice, in: Proceedings of the Third Workshop on Computational Approaches to Linguistic Code-Switching, 2018, pp. 82–91.

[244] E. Ahn, C. Jimenez, Y. Tsvetkov, A. Black, What code-switching strategies are effective in dialogue systems?, Proceedings of the Society for Computation in Linguistics 3 (1) (2020) 308–318.

[245] R. Reddy, N. Reddy, S. Bandyopadhyay, Dialogue based Question Answering System in Telugu, in: Proceedings of the Workshop on Multilingual Question Answering-MLQA, 2006.

[246] S. Banerjee, N. Moghe, S. Arora, M. M. Khapra, A dataset for building code-mixed goal oriented conversation systems, COLING.

[247] V. Ramanarayanan, R. Pugh, Y. Qian, D. Suendermann-Oeft, Automatic turn-level language identification for code-switched spanish–english dialog, in: 9th International Workshop on Spoken Dialogue System Technology, Springer, 2019, pp. 51–61.

[248] A. Srivastava, K. Bali, M. Choudhury, Understanding script-mixing: A case study of hindi-english bilingual twitter users, in: Proceedings of the The 4th Workshop on Computational Approaches to Code Switching, 2020, pp. 36–44.

[249] D. Garrette, H. Alpert-Abrams, T. Berg-Kirkpatrick, D. Klein, Unsupervised code-switching for multilingual historical document transcription,

in: Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2015, pp. 1036–1041.

[250] L. Qin, M. Ni, Y. Zhang, W. Che, Cosda-ml: Multi-lingual code-switching data augmentation for zero-shot cross-lingual nlp, arXiv preprint arXiv:2006.06402.

[251] A. R. KhudaBukhsh, S. Palakodety, J. G. Carbonell, Harnessing code switching to transcend the linguistic barrier, arXiv preprint arXiv:2001.11258.

[252] J. Yang, S. Ma, D. Zhang, S. Wu, Z. Li, M. Zhou, Alternating language modeling for cross-lingual pre-training., in: AAAI, 2020, pp. 9386–9393.

[253] A. Conneau, G. Lample, R. Rinott, A. Williams, S. R. Bowman, H. Schwenk, V. Stoyanov, Xnli: Evaluating cross-lingual sentence representations, arXiv preprint arXiv:1809.05053.

[254] M. Diab, J. Hirschberg, P. Fung, T. Solorio, Proceedings of the first workshop on computational approaches to code switching, in: Proceedings of the First Workshop on Computational Approaches to Code Switching, 2014.

[255] G. Molina, F. AlGhamdi, M. Ghoneim, A. Hawwari, N. Rey-Villamizar, M. Diab, T. Solorio, Overview for the second shared task on language identification in code-switched data, arXiv preprint arXiv:1909.13016.

[256] R. S. Roy, M. Choudhury, P. Majumder, K. Agarwal, Overview of the fire 2013 track on transliterated search, in: Post-Proceedings of the 4th and 5th Workshops of the Forum for Information Retrieval Evaluation, 2013, pp. 1–7.

[257] P. R. Rao, S. L. Devi, Cmee-il: Code mix entity extraction in indian languages from social media text@ fire 2016-an overview., in: FIRE (Working Notes), 2016, pp. 289–295.

[258] R. Sequiera, M. Choudhury, P. Gupta, P. Rosso, S. Kumar, S. Banerjee, S. K. Naskar, S. Bandyopadhyay, G. Chittaranjan, A. Das, et al., Overview of fire-2015 shared task on mixed script information retrieval., in: FIRE Workshops, Vol. 1587, 2015, pp. 19–25.

[259] S. Banerjee, K. Chakma, S. K. Naskar, A. Das, P. Rosso, S. Bandyopadhyay, M. Choudhury, Overview of the mixed script information retrieval (msir) at fire-2016, in: Forum for Information Retrieval Evaluation, Springer, 2016, pp. 39–49.

[260] A. Jamatia, A. Das, Task report: Tool contest on pos tagging for code-mixed indian social media (facebook, twitter, and whatsapp) text@ icon 2016., Proceedings of ICON.

[261] K. Prahallad, A. Vadapalli, S. Kesiraju, H. Murthy, S. Lata, T. Nagarajan, M. Prasanna, H. Patil, A. Sao, S. King, et al., The blizzard challenge 2014, in: Proc. Blizzard Challenge workshop, Vol. 2014, 2014.

[262] X. Shi, Q. Feng, L. Xie, The asru 2019 mandarin-english code-switching speech recognition challenge: Open datasets, tracks, methods and results, arXiv preprint arXiv:2007.05916.

[263] S. Khanuja, S. Dandapat, A. Srinivasan, S. Sitaram, M. Choudhury, Gluecos: An evaluation benchmark for code-switched nlp, arXiv preprint arXiv:2004.12376.

[264] G. Aguilar, S. Kar, T. Solorio, Lince: A centralized benchmark for linguistic code-switching evaluation, in: Proceedings of The 12th Language Resources and Evaluation Conference, 2020, pp. 1803–1813.