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 list what is available in various code-switched language pairs with the language processing tasks they can be used for. We review code-switching research in various Speech and NLP applications, including language processing tools and end-to-end systems. 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 settings such as navigation instructions and formal settings such as news paper headlines and formal 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 start permeating more and more applications that involve interactions with humans [8, 9], it is imperative that they take phenomena such as code-switching into account.

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 example, 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 of course 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 conversations, we need to also be able to understand what is being said in these 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. 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. 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 [10, 11, 12], 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.

We would like to put this work in a common framework, but its probably still too early to fix such a framework. 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 simple language technology tasks that can be achieved with simple analysis, 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 [13], that there is an underlying language choice which mostly defines the grammar and morphological aspects of the utterance, but it is also clear that the matrix language can change during 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 going to be hard, 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 theoreti-

cal aspects. In Section 3 we describe speech and NLP corpora and resources that have been created for code-switched language pairs. Section 4 describes approaches to handling code-switching in specific speech and NLP applications. We conclude in Section 5 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. As mentioned earlier, code-switching is not a random juxtaposition of the two languages being mixed. 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 code-switching has been studied by linguists for decades and providing a complete description of their research 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'[14], while code-mixing is 'the embedding of linguistic units such as phrases, words and morphemes of one language into an utterance of another language'[15]. The distinction between code-switching, mixing and lexical borrowing is often not clear and can be thought of as lying on a continuum [16]. 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. [17] 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 [18]. 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 [19] 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.

[20] 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. [21] showed that there exists a constituent tree labeling, implying that around a switch point there is a constraint on an equivalence order in constituents. The above described linguistic theories are also used in [22] to identify governing relationships between constituents. [23] have demonstrated evidence that a constrained Universal Grammar needs refinement of f-selection in code-switching as compared to monolingual speech. [24] have proposed four categories for any switch point comprising of harmonization,

neutralization, compromise, and blocking.

[25] 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 [26]. 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 [27] 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 [28]. 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. [29] have studied attested contact-induced changes based on prior linguistic theories regarding the types of structural changes in calques, distributions, frequencies, inventory and stability.

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) [30] is an utterance and corpus level

metric proposed to measure the amount of code-switching in corpora by using word frequencies. [31] 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. [32] 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.

[32] 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 [33].

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 [34] 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.
- The HKUST Mandarin-English Corpus [35] also consists of interviews and conversational speech with 5 hours of transcribed and 15 hours of untranscribed speech.
- The CECOS corpus [36] contains 12 hours of prompted Mandarin-English speech from 77 speakers.
- The CUMIX Cantonese-English speech corpus [37] contains 17 hours of code-switched speech read by 80 speakers.
- A Mandarin-Taiwanese corpus is described in [38] containing 4000 utterances recorded by 16 speakers.
- BANGOR-MIAMI [39] 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 [40].

- [41] 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 [42] consists of 50 hours of conversational speech between Hindi-English bilinguals. There are around 500 speakers in the corpus.

- [43] 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.

- [44] 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.

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

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

- An Egyptian Arabic-English speech corpus is described in [47]. 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) [48] 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)[49] consists of 7.3 hours of French-Algerian Arabic code-switched speech from 20 bilingual speakers.

- [50] 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 [51]. It contains around 16 hours of transcribed speech with code-switching boundary annotations. This corpus is extended in [52] 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 [53]. 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.

- 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 [54].

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. 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 [55]. 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 [55], another effort was made on a larger scale to collect 5933 questions for Hindi-English, Tamil-English, Telugu-English grounded on articles and images [56].

- 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 [57].

- 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 [58].

3.2.2. 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 [59] [60].

- A predetermined set of 11 users of Facebook users were selected to search for publicly available content that resulted in 2335 posts and 9813 com-

ments [61]. 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 [62] to collect more data from Bengali, Hindi and English with finer annotation schema catering to named entities and also explicitly annotating suffixes.

- [63] 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.

- Another very large scale dataset that is is not explicitly targeted at code-switching but contains it is [64] that addresses curating socially representative text by caring about geographic, social, topical and multilingual diversity. This corpus consists of Tweets from 197 countries in 53 languages.

3.2.3. 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 [65].

- Twitter is a commonly used source of code-switched data. [66] annotated 3,638 tweets with three Named Entity tags 'Person', 'Organization' and 'Location' using the BIO scheme.

3.2.4. 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 [67].
- 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 [68].
- 922 sentences of spoken Spanish-English conversational data is transcribed and annotated with POS tags in [69].
- [70] gathered 1106 messages (552 Facebook posts and 554 tweets) in Hindi-English and annotated them with a Twitter specific tagset. [71] 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.
- [72] 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.

3.2.5. Parsing

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

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

- [74] 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 [75].
- Code-switched test utterances for the NLmaps corpus are constructed by [76]. 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.

3.3. Information retrieval

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

4. Code-switched Speech and NLP

[**?**] present a brief survey of code-switching studies in NLP. 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 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 [78], however, in practice such resources are not available and other techniques such as synthesizing code-switched data for training such embeddings can be used.

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 [79]. Another approach runs multiple recognizers in parallel with an LID system and uses scores from all the systems for decoding speech [80]. In [46], no LID system is used - instead, two recognizers in English and Malay are run in parallel and the hypotheses produced 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. [38] 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. [81] develop a cross-lingual phonetic Acoustic Model for Cantonese-English speech, with the phone set designed based on linguistic knowledge. [82] 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 com-

bination of the phone inventories. [42] 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. [83] 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 [84] 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.

[53] 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. [44] 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. [85] use cross-lingual data sharing to tackle the problem of highly imbalanced Mandarin-English code-switching, where the speakers speak primarily in Mandarin.

When data from both languages is available but there is no or very little data in code-switched form, bilingual models can be built. [83] train the Acoustic Model on bilingual data, while [86] and [87] use existing monolingual models and with a phone-mapped lexicon and modified Language Model for Hindi-English ASR.

[88] 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. [89] decode untranscribed data with this ASR system and add the decoded speech to ASR training data after rescoring using Language Models. In [90], 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.

[52] build a unified ASR system for five South African languages, by using 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.

[91] 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.

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. [92] 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. [93] 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.

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. [94] 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 [95] 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. [96] use inversion constraints to predict CS points and integrate this prediction into the ASR decoding process. [97] 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. [98] 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.

[99] 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. [78] 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.

[**?**] 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 [100], [**?**] and [101], RNNLMs are combined with n-gram based models, or converted to backoff models, giving improvements in perplexity and mixed error rate. [102] 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.

4.3. code-switching detection from 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.

[103] 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.

4.4. Speech Synthesis

Currently, 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, phenomenon 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 [104, 105, 106]. Other approaches utilise cross-language voice conversion techniques [107] and adaptation on a combination of data from multiple languages[108]. Assimilation at the linguistic level is fairly successful for phonetically similar languages [106], 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.

[109] 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. [110, 111] propose a method to use a word to phone mapping instead, where an English word is statistically mapped to Indian language phones.

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. A large amount of research in this area has been conducted due to shared tasks on word-level LID ([59], [60]). Social media data, especially posts from Facebook was used to collect data for the task of LID [61] 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.

[112] 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. [113] make use of patterns in language usage of Hinglish along with the consecutive POS tags for LID. [62] have also experimented with n-gram modeling with pruning and SVM based models with feature ablations Hindi-English and Bengali-English LID. [63] have worked on re-defining and re-annotating language tags from social media cues based on cultural, core and therapeutic borrowings. [64] 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. [114] 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.

4.6. Named Entity Recognition

Another sequence labeling task of interest is Named Entity Recognition (NER). [65] organized a shared task on NER in code-switching by collecting data from tweets for Spanish-English and Arabic-English. [115] 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. [116] 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, [117] 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.

[118] studied Arabic text on social media by exploring the influence of word embedding based representations on NER. Along similar lines, [119] also investigated how word representations are capable of boosting semi-supervised approaches to NER. [66] 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 [120].

4.7. POS Tagging

Recently there has been interest in code-switched structured prediction tasks like POS tagging and parsing. [67] 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. [68] presented POS annotation for Turkish-German tweets that align with existing language identification based on POS tags from Universal Dependencies. [69] 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.

[70] 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. [72] 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.

4.8. Parsing

[121] 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 WordNet through a Word Sense Disambiguator [122]. 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.

[74] 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. [75] 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.

[76] performed multilingual semantic parsing using a transfer learning approach for code-switched text utilizing cross lingual word embeddings in a sequence to sequence framework. [73] compared different systems for dependency parsing and concluded that the Multilingual BIST parser is able to parse code-switched data relatively well.

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 [55]. 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 [123]. 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, [56] 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 [124] which also has an additional advantage of collecting more data.

[57] 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. [58] 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

[125].

[126] 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 [127], [128].

4.10. Machine Translation

[131] developed a machine translation scheme for translating Hinglish into pure English and pure Hindi forms by performing cross morphological analysis. [132] 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.

4.11. Dialogue and discourse

[133] 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.

[134] 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.

Cross-lingual Question Answering systems were extended to dialog systems for railway inquiries [129]. Recently, there has been an attempt to create code-mixed version of goal oriented conversations [130] from the DSTC2 restaurant reservation dataset.

It is clear from the description of the tools and language processing applications above that 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. 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 psycholinguistics to understand how, when and why to code-switch.

5. Challenges and Future Directions

Although code-switching is a persistent phenomenon through out the whole world, access to data will always be hard. 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. Also as code-switching is more likely to be used in less task specific contexts, with less explicit function it may also be harder to label such 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.

Also most code-switching studies focus on pairs with one high resource language (e.g. English, Spanish, MSA, Putonghua) and a lower resource language, but 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.

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.

We are not pretending that language technologies for code-switching is 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. 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. Such systems will come, and they will use the results of the work surveyed here.

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 developing techniques that can work over multiple pairs of code-switched languages may lead to faster development and generalization of the field.

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

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

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

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

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

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

[16] 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.

[17] 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.

[18] 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.

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

[20] 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.

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

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

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

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

[25] 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.

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

[27] 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.

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

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

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

[31] 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.

[32] 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.

[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), 2017.

[34] 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.

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

[36] H.-P. Shen, C.-H. Wu, Y.-T. Yang, C.-S. Hsu, Cecos: A Chinese-English Code-Switching Speech Database, in: Speech Database and Assessments (Oriental COCOSDA), 2011 International Conference on, IEEE, 2011, pp. 120–123.

[37] 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.

[38] 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.

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

[40] 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.

[41] 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.

[42] 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.

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

[44] 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.

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

[46] 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.

[47] 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.

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

[49] 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.

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

[51] 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.

[52] 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.

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

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

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

[56] 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.

[57] 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.

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

[59] 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.

[60] 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.

[61] 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.

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

[63] 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.

[64] 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.

[65] 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.

[66] 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.

[67] 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.

[68] 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.

[69] 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.

[70] 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.

[71] 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.

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

[73] 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.

[74] 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.

[75] 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.

[76] 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.

[77] 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.

[78] 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.

[79] 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.

[80] 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.

[81] 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.

[82] 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.

[83] 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.

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

[85] 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.

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

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

[88] 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.

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

[90] 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.

[91] 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.

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

[93] 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.

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

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

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

[97] 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.

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

[99] 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.

[100] 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.

[101] 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.

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

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

[104] 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.

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

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

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

[108] 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.

[109] 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.

[110] 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.

[111] 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.

[112] 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.

[113] 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.

[114] 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.

[115] 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.

[116] 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.

[117] 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.

[118] 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.

[119] 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.

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

[121] 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.

[122] 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.

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

[124] 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.

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

[126] 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.

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

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

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

[130] S. Banerjee, N. Moghe, S. Arora, M. M. Khapra, A dataset for building code-mixed goal oriented conversation systems, arXiv preprint arXiv:1806.05997.

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

[132] 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.

[133] 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.

[134] 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.