Multilingual Identification of English Code-Switching

Igor Sterner
Department of Computer Science and Technology
University of Cambridge, UK
is473@cam.ac.uk

Abstract

Code-switching research depends on fine-grained language identification. In this work, we study existing corpora used to train token-level language identification systems. We aggregate these corpora with a consistent labelling scheme and train a system to identify English code-switching in multilingual text. We show that the system identifies code-switching in unseen language pairs with absolute F_1 measure 2.3-4.6% better than language-pair-specific SoTA. We also analyse the correlation between typological similarity of the languages and difficulty in recognizing code-switching.

1 Introduction

Code-switching is when bilinguals alternate between languages at the sentence or word level. Increasing attention is being placed on computational approaches to code-switching, driven by six code-switching workshops to date (Solorio et al., 2014; Molina et al., 2016; Aguilar et al., 2018b; Solorio et al., 2020, 2021; Winata et al., 2023). In part, this line of research is due to the rise in the use of code-switching on social media (Jose et al., 2020), potentially as a result of language contact (Gardner-Chloros, 2020).

Language technology users now expect automatic speech recognition systems, text-to-speech engines, generative models etc. to handle code-switching as a natural form of language. But even SoTA large language models (LLMs) perform poorly on zero-shot NLP tasks with code-switching data (Zhang et al., 2023). They are outperformed by smaller fine-tuned models. Further, Yong et al. (2023) report acceptability judgements of LLM-generated code-switching, showing few generations are acceptable. Despite the prevalence of code-switching in spoken and online discourse, code-switching is likely a linguistic phenomenon severely underrepresented in the training data of models like those in the GPT family (Brown et al., 2020). The availability of code-switching data has therefore become a common barrier to address the limitations of existing NLP tools on code-switching input. A tool required to address this barrier is fine-grained and multilingual language identification systems.

In this paper, we develop a fine-grained tool that distinguishes words between English and any other language.[1] We make our models and code available.[2]

2 Background

There are many works aimed at identifying languages in documents at more fine-grained levels, e.g. the word-level (Lyu and Lyu, 2008; Solorio et al., 2014; Mave et al., 2018; Zhang et al., 2018; Nguyen et al., 2021; Hidayatullah et al., 2022; Hegde et al., 2024) or even sub-word level (Mager et al., 2019; Sabty et al., 2021). Figure 1b shows the annotation scheme of one German–English work which aims for very fine-grained classification.

Approaches to compile code-switching corpora traditionally involved collecting spoken recordings of bilinguals (Myers-Scotton, 1992; Deuchar, 2009; Nguyen and Bryant, 2020) or more recently generating synthetic code-switching (Chang et al., 2019; Gupta et al., 2020; Rizvi et al., 2021). Manually collecting recordings is an expensive, arduous and lengthy task; meanwhile, synthetic code-switching is inherently limited in the code-switching phenomena it exhibits. But with automatic code-switching identification systems, much larger corpora of naturally occurring code-switching have begun to be collected (Nayak and Joshi, 2022; Sterner and

[1] In practice, the choice to focus only on languages switched with English was as a result of data availability.

[2] Code-switching identification: https://huggingface.co/igorsterner/AnE-LID, binary named entity recognition: https://huggingface.co/igorsterner/AnE-NER, code: https://github.com/igorsterner/AnE

Proceedings of the Eleventh Workshop on NLP for Similar Languages, Varieties, and Dialects (VarDial 2024), pages 163–173

Teufel, 2023; Wintner et al., 2023). The source of such data is large troves of social media posts.

Such corpora offer the potential to test various theories of code-switching, theories of when humans code-switch and why. An example is the triggering hypothesis (Clyne, 1980), which suggests that shared lexical items (e.g. named entities) are triggers of code-switching. Broersma and De Bot (2006) and Broersma (2009) found statistical significance between such lexical triggers and code-switching points for a small handful (c. 100) of switch points in recorded corpora. Soto et al. (2018) test on the larger Spanish–English spoken corpus of Deuchar (2009), but limit their study to a small list of cognates. Wintner et al. (2023) test on data of three different language pairs (Arabic–English, Spanish–English and German –English) with a total of 648,498 switch points from almost 10M tokens of mostly automatically language-identified social media content. They found statistical correlation suggesting switch points tend to be close to the shared lexical items. For this to be possible, substantial effort was invested to build word-level language identifiers specific to each of the three language pairs they explore (Aguilar et al., 2020; Shehadi and Wintner, 2022a; Osmelak and Wintner, 2023). Corpus-linguistic approaches to code-switching will continue to depend on the quality of such fine-grained language identification tools.

Existing code-switching identification systems are language-specific; they distinguish between a fixed number of (typically two) specified languages in a training corpus. This approach fails to support code-switching research in lower-resource languages, where annotated training data is either not available or available at much smaller scales. To collect more data for low-resource language pairs requires an identification system, a circular problem. This circular problem applies more generally to language identification. But it is especially challenging in code-switching because code-switching sentences are often only found in seas of spoken/written data of primarily monolingual sentences.

3 Existing Corpora

Large language-identified corpora of code-switching with English only exist in a small set of language pairs, namely Hindi–English (Singh et al., 2018), Spanish–English (Molina et al., 2016; Aguilar et al., 2018a), Nepali–English (Solorio et al., 2014), German–English (Osmelak and Wintner, 2023) and Arabic–English (Shehadi and Wintner, 2022b). Smaller corpora of code-switching of low-resource language pairs also exist, e.g. Indonesian–English (Barik et al., 2019), Turkish–English (Yirmibeşoğlu and Eryiğit, 2018) and Vietnamese–English (Nguyen and Bryant, 2020). These corpora are derived from posts on social media platforms such as Twitter and Reddit, except for the Vietnamese–English corpus which is of spoken code-switching.

Of the corpora, there is a variation in the labelset used to classify the words. The variation is centred around the annotation of shared words and words of mixed morphology. Example labelsets, alongside the frequency of words of each label, are given in Tables 1 and 2.

In many public corpora of low-resource language pairs, code-switching is identified at a coarser-grained level. These corpora only include labels for each of the two languages, and sometimes a third label for all tokens not of the two languages. Meanwhile, higher-resource language pairs include the identification of named entities, or more generally shared words, mixed words and foreign words not of either the two languages in question. The labelset proposed by (Molina et al., 2016, the second shared task on language identification in code-switching) includes ‘lang1’, ‘lang2’, ‘other’, ‘ne’ (named entity), ‘fw’ (foreign word), ‘mixed’, ‘unk’ and ‘ambiguous’ labels. This labelset was adapted from Solorio et al. (2014, the first shared task) which has the same labels except without ‘fw’ or ‘unk’.

Hindi–English, Spanish–English and Nepali–English code-switching datasets have been brought together in the LinCE benchmark (Aguilar et al., 2020), under a language identification (LID) task for code-switching data. They use the labels of Molina et al. (2016) or Solorio et al. (2014).

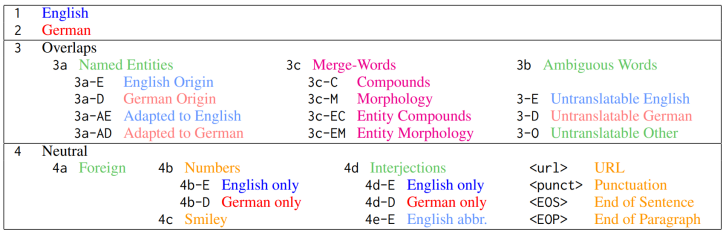
In addition to code-switching identification, LinCE includes a benchmark for named entity recognition (NER) in code-switching data. The code-switching examples in the LID and NER benchmarks are different.

In the Denglisch corpus of Osmelak and Wintner (2023), German–English code-switching is identified at a more fine-grained level. Figure 1b displays the fine-grained labels they annotate words for, demonstrating the number of linguistic phenomena in code-switching inter-play. In their work, they use 100% of their human-annotated data in the

	1	2	punct	EOS	EOP	4b	3a	3a-D	3a-E	3a-AD	4a	3a-AE	url	4c	3-O	3c-C
Train (3364)	24134	23598	9016	3351	2976	899	460	405	311	210	206	191	184	99	96	80
Dev (420)	2621	3093	992	420	212	58	51	37	46	26	18	18	15	10	2	11
Test (421)	3125	2914	1082	417	279	80	62	43	35	26	19	21	10	9	12	9

	3c-M	4d	4b-D	3-D	3b	4	3-E	4d-D	3c-EC	4e-E	4b-E	4d-E	3c	3	3c-EM
Train (3364)	76	71	65	60	58	48	46	41	22	15	14	14	12	7	5
Dev (420)	13	10	4	0	13	1	9	7	1	0	3	5	2	0	1
Test (421)	9	7	5	0	16	2	6	3	5	1	0	2	0	0	0

(a) Label frequencies



(b) Annotation scheme. Source: Osmelak and Wintner (2023)

Table 1: Details of the Denglisch corpus of German–English code-switching (Osmelak and Wintner, 2023)

	lang1	lang2	other	ne	fw	mix	unk	amb
Train (4823)	54720	19134	14017	6069	398	33	10	8
Dev (744)	8942	3303	2210	837	29	5	2	1
Test (1854)	20635	7487	5369	2432	106	14	5	32

(a) Hindi–English (Singh et al., 2018)

	lang2	lang1	other	ne	amb	unk	mix	fw
Train (21030)	111422	77843	53851	4725	263	210	27	22
Dev (3332)	14787	16618	7810	769	37	32	3	2
Test (8289)	42850	31916	20311	2059	100	80	17	8

(b) Spanish–English (Molina et al., 2016)

	lang2	lang1	other	ne	mix	amb
Train (8451)	49936	38827	29847	3146	90	72
Dev (1332)	8385	5557	4653	452	13	11
Test (3228)	19881	14009	11321	1268	48	32

(c) Nepali–English (Solorio et al., 2014)

	id	un	en
Test (825)	11200	5917	5608

(d) Indonesian–English (Barik et al., 2019)

	t	e
Test (377)	3941	1489

(e) Turkish–English (Yirmibeşoğlu and Eryiğit, 2018)

	@vie	@eng	@non
Test (3313)	16974	7219	614

(f) Vietnamese–English (Nguyen and Bryant, 2020)

Table 2: Code-switching identification corpora, with frequencies of labels. lang1 is always English.

cross-validation setup. Their data can be collapsed to have a labelset similar to the data in LinCE.

The Arabic–English code-switching dataset contains labels for 'Shared Other' words, which are less simple to adapt to the LinCE labelset, likely requiring some further annotation.

For low-resource language pairs, Turkish–English (Yirmibeşoğlu and Eryiğit, 2018) includes only binary labels (Turkish and English), Indonesian–English (Barik et al., 2019) adds an 'other' (or 'unknown' as they called it) category for named entities, punctuation and other non-language units. The Vietnamese–English CanVEC corpus includes the same three categories, but their data is semi-automatically annotated; a human only corrects words not contained in wordlists of either language, and words in both wordlists.

SoTA language identification performance for the high-resource language pairs is displayed on the LinCE benchmark leaderboard.[3] As of 23 April 2024, the best system is the XLM-RoBERTa language model (Conneau et al., 2020) fine-tuned separately for classification on each of the language pairs. This is an anonymous submission and no reference is given to the exact training setup. There is no existing language identification baseline on the Vietnamese–English corpus, likely because it is semi-automatically annotated data. For Indonesian–English and Turkish–English, SoTA language identification performance remains from the original works; both using conditional random field (CRF) classifiers. Like Osmelak and Wintner (2023) do for German–English, these systems also use 100% of their corpora in the cross-validation setup. They release no separate test set.

The disparity in size and labelset of these code-switching corpora has presented a challenge to research in this field. The best code-switching iden-

[3] https://ritual.uh.edu/lince/leaderboard

tification systems are language pair-specific. This has left low-resource language pairs behind in code-switching research. In addition, there is no baseline for research on new language pairs to evaluate against.[4]

4 AnE

Our goal is to develop an Any-English (AnE) code-switching identification system, which we reformulate as the task of identifying English code-switching in a sea of text of other languages. English here encompasses the many local varieties of English present in the aforementioned corpora of code-switching. We aim to achieve our goal by matching up the labelsets of existing corpora with this task in mind. A key challenge we face is that some corpora distinguish named entities, whilst others do not.

To alleviate this challenge, we will train two classifiers:

1. **Code-switching identification -** one will distinguish between *English*, other languages (hereinafter *notEnglish*), words that mix English and another language within the word (*Mixed*) and other words such as punctuation, emojis and mentions (*Other*).
2. **Binary named entity recognition -** the other will make a binary distinction as to whether a word is part of a named entity or not.

The data we searched for to train these classifiers broadly fits into three categories. The first is corpora that classify the language of the words but also have a named entity class (LID+NER). These corpora are labelled with the previously discussed labelsets of Molina et al. (2016) or Solorio et al. (2014). The second is corpora that only classify the language of the words as L1 or L2 (LID). Some of these corpora also have an 'other' category which includes named entities/punctuation/emojis etc., and some simply remove 'other' words from the data by manual means. The third is derived from the task of named entity recognition on code-switching text; such corpora include the named entity labels and classes in BIO (Ramshaw and Marcus, 1995) format (NER).

We preprocess the corpora as follows.

[4]Except by prompting LLMs, of which only the largest models perform well (Zhang et al., 2023). This is currently a subpar and prohibitively expensive solution.

	English	notEnglish	Mixed	Other
Train (4823)	54720	19550	33	14017

(a) Hindi–English (LinCE-LID, Singh et al., 2018)

Train (21030)	77843	111917	27	53851

(b) Spanish–English (LinCE-LID, Molina et al., 2016)

Train (33611)	78588	199723	45	110015

(c) Spanish–English (LinCE-NER, Aguilar et al., 2018a)

Train (8451)	38827	50008	90	29847

(d) Nepali–English (LinCE-LID, Solorio et al., 2014)

Train (3364)	24725	24865	195	16525

(e) German–English (Osmelak and Wintner, 2023)

Table 3: Collapsed LID training data statistics

	I	O
Train (4823)	6069	88320

(a) Hindi–English (Singh et al., 2018)

Train (21030)	4725	243638

(b) Spanish–English (Molina et al., 2016)

Train (8451)	3146	118772

(c) Nepali–English (Solorio et al., 2014)

Train (3364)	1577	65193

(d) German–English (Osmelak and Wintner, 2023)

Train (1243)	2222	17806

(e) Hindi–English (Singh et al., 2018)

Train (33611)	11722	385055

(f) Spanish–English (Aguilar et al., 2018a)

Table 4: Binary NER training data statistics

- **LID+NER** Each corpora becomes two subcorpora. In the first, the language other than English, foreign words, ambiguous words and unknown words all become *nonEnglish*. The English, Mixed and Other tags stay as *English*, *Mixed* and *Other*. Named entities receive a special ID to be ignored in all training updates. In the second subcorpora, named entities become a generic inside (*I*) and all other labels become outside (*O*).
- **LID** All labels are taken directly, which always includes *English* and *notEnglish*. *Other* is also taken if included in the data. There were no *Mixed* labels in any of these corpora.

- **NER** All B or I labels of any type become an inside (*I*) label. All outside (*O*) labels stay.

Table 3 gives statistics for the output from the collapse of the corpora into our LID scheme, and Table 4 for the collapse into binary NER. Statistics for the LID-only category of corpora follow directly from Table 2 (d)-(f).

5 Experiment

5.1 Experimental Setup

Systems and Baselines Our AnE system is an ensemble of the language identification (AnE_{LID}) and binary named entity recognition (AnE_{NER}) classifiers. The ensemble is achieved by classifying words based on each classifier separately, and then overwriting labels of words AnE_{NER} predicts to be named entities. For the high-resource language pairs, these labels make up the *NamedEntity* class. For the low-resource language pairs with an *Other* category, they are moved into there. For languages without any *Other* category, AnE_{NER} is not used and AnE_{LID}'s *Other* labels become *NotEnglish*. The low-resource language pair copora also do not include a *Mixed* category, so such predicted words become *NotEnglish*.

For the LinCE benchmark language pairs, and German–English, we train separate baseline classifiers using only data from each single language pair. This baseline corresponds to reproducing the SoTA (anonymous) system on the LinCE leaderboard. For the low-resource language pairs, we test on 100% of the data. Therefore, we are unable to train a baseline system. Instead, we use the best-performing system from the original works as baseline, even though these were trained in the cross-validation setup.[5]

All classifiers are single-layer perceptron classification heads on XLM-RoBERTa (large).

Data We use the data described in Section 4. In particular, we use the provided splits from the LinCE benchmark (Aguilar et al., 2020), which includes three language pairs. We also mix in Denglisch (German–English) data from Osmelak and Wintner (2023). In their work, they train with the cross-validation setup. We instead split their data into train/dev/test with splits 80:10:10%.

We balance the training data between these four language pairs by up-sampling until all language pairs contribute the same number of training sentences.

We also evaluate on 100% of the three low-resource language pair corpora, namely Indonesian–English, Turkish–English and Vietnamese–English. We remind the reader that the Vietnamese–English corpus is different to all other corpora in that (a) it is a corpus of spoken code-switching and (b) it is only silver-standard data.

Training We train all systems for 3 epochs with a learning rate of 1e-5 and a batch size of 32. All parameters are updated using a cross-entropy loss criterion and the Adam optimizer (Kingma and Ba, 2014). We use weight decay = 0.01 for the optimizer with $\beta = (0.9, 0.999)$ and ϵ =1e-8. For the named entity tokens without language subcategorization, as described in Section 4, losses are zeroed. These hyperparameters were chosen based on recommendations from prior work (e.g., Devlin et al., 2019). No hyperparameter tuning was performed.

When training the baseline systems, we continue training for additional epochs until the same number of sentences are seen as in the up-sampled AnE data for the language pair in question. We found validation accuracy monotonically increases and plateaus by the end of training; there was no evidence of overfitting despite this extended training setup.

Metrics We will compare the performance of the AnE system against baseline by computing precision (P), recall (R) and weighted-average F_1 measure. All the measures are word-based. The LinCE submission portal generates P, R and F_1 metrics for each label, and an overall weighted F_1 measure.[6] We also use weighted-average F_1 measure for other evaluations.

XLM-RoBERTa uses byte-pair encoding for subword tokenization. If there is more than one unique subword label for a given word, we select the most frequent label. In the event of a tie, we select the label which appeared first. This detail is likely to particularly affect the classification of mixed-morphology words, which will often be split into subwords. Further investigation of this effect is beyond the scope of this work.

[5] Therefore, numbers are not directly comparable. But either way, our system is not favoured as it does not have any training data for these language pairs.

[6] No overall P or R is provided.

	English			notEnglish			Mixed		
	P	R	F_t	P	R	F_t	P	R	F_t
Hindi–English		*(20635)*			*(7487)*			*(14)*	
hi-en only	98.39	98.47	98.43	95.77	96.66	96.21	47.37	64.29	54.55
AnE	98.32	98.49	98.40	94.24	96.61	95.41	61.54	57.14	59.26
Spanish–English		*(42850)*			*(31916)*			*(17)*	
es-en only	98.22	98.94	98.58	98.98	99.22	99.10	0.00	0.00	0.00
AnE	98.51	98.62	98.57	99.01	99.13	99.07	54.55	35.29	42.86
Nepali–English		*(19881)*			*(14009)*			*(48)*	
ne-en only	96.34	96.90	96.62	98.27	98.07	98.17	54.29	39.58	45.78
AnE	96.71	96.25	96.48	97.78	98.40	98.09	62.50	41.67	50.00
German–English		*(3134)*			*(2978)*			*(23)*	
de-en only	98.76	99.23	99.00	99.39	98.93	99.16	78.26	78.26	78.26
AnE	97.13	99.27	98.19	99.06	98.62	98.84	63.64	60.87	62.22

	Named Entity			Other			Ambiguous		
Hindi–English		*(2432)*			*(5369)*			*(32)*	
hi-en only	90.18	89.14	89.66	99.16	98.96	99.06	0.00	0.00	0.00
AnE	91.77	87.54	89.60	98.38	98.44	98.41	0.00	0.00	0.00
Spanish–English		*(2059)*			*(20311)*			*(100)*	
es-en only	87.76	82.18	84.88	99.82	99.78	99.80	0.00	0.00	0.00
AnE	77.45	81.40	79.37	99.81	99.82	99.82	0.00	0.00	0.00
Nepali–English		*(1268)*			*(11321)*			*(32)*	
ne-en only	73.07	74.68	73.87	97.63	97.32	97.48	8.33	3.12	4.55
AnE	72.52	75.95	74.19	97.68	97.06	97.37	0.00	0.00	0.00
German–English		*(187)*			*(1877)*				
de-en only	90.67	93.58	92.11	100.00	99.63	99.81	-	-	-
AnE	88.54	90.91	89.71	99.89	96.70	98.27	-	-	-

	Unknown			Foreign Word			Overall		
Hindi–English		*(5)*			*(106)*			*(36080)*	
hi-en only	0.00	0.00	0.00	87.10	50.94	64.29	-	-	97.33
AnE	0.00	0.00	0.00	0.00	0.00	0.00	-	-	96.86
Spanish–English		*(80)*			*(8)*			*(97341)*	
es-en only	50.00	5.00	9.09	0.00	0.00	0.00	-	-	98.58
AnE	0.00	0.00	0.00	0.00	0.00	0.00	-	-	98.44
Nepali–English								*(46559)*	
ne-en only	-	-	-	-	-	-	-	-	96.76
AnE	-	-	-	-	-	-	-	-	96.66
German–English								*(8199)*	
de-en only	-	-	-	-	-	-	99.03	99.02	99.03
AnE	-	-	-	-	-	-	98.17	98.15	98.15

Table 5: Results for the LID task for language pairs in the training data

5.2 Results

Table 5 gives test results on the four language pairs included in the training data. Mixing the data to train one AnE model does not result in a large change in performance compared to the separate baseline models. Overall F_1 measures for the baseline and AnE are 97.33/96.86% for Hindi–English, 98.58/98.44% for Spanish–English, 96.76/96.66% for Nepali–English, and 99.03/98.15% for German–English. AnE is numerically worse for all language pairs, but only by a small margin of less than 1% absolute F_1.

For the first three, which are all from LinCE, the differences are all less than 0.5%. For German–English, it is slightly larger (-0.88%). We collapsed the labels for German–English to match the LinCE evaluation labels where possible. But there may be some differences between the LinCE data and the German–English data scheme. This may be a cause of the slightly greater drop in the overall performance of AnE for this language pair.

AnE also does not have predictive classes 'Ambiguous', 'Unknown' or 'Foreign Word'. There are few (all < 106) words in these categories in the test data. Nevertheless, the AnE system scores zero for all these categories, which may be another reason for the small numerical drop in overall performance compared to the baselines.

The separate baseline classifiers perform near-identical to the anonymous SoTA reported on the LinCE leaderboard.

In terms of evaluating our approach of separating out the binary NER task, the results show that AnE_{NER} in the AnE ensemble is near-identical to the baseline where 'Named Entity' is simply a label amongst the other labels. In particular, named entity F_1 measure for the baseline and AnE is recorded at 89.66/89.60% for Hindi–English, 84.88/79.37% for Spanish–English, 73.87/74.19% for Nepali–English and 92.11/89.71% for German–English. The reduced performance in Spanish–English compared to the baseline can be attributed to a substantially worse precision (77.45 vs. the baseline 87.76). AnE_{NER} was trained on both the

	English			notEnglish			Other			Overall		
	P	R	F_t	P	R	F_t	P	R	F_t	P	R	F_t
Indonesian–English		*(5608)*			*(11200)*			*(5917)*			*(22725)*	
id-en SoTA	89.90	84.42	87.07	88.13	96.22	91.99	94.99	83.96	89.14	90.70	87.38	88.86
AnE	86.86	97.63	91.93	95.44	94.86	95.15	97.13	86.83	91.69	93.76	93.45	93.45
Turkish–English		*(1489)*			*(3941)*						*(5430)*	
tr-en SoTA	91.7	92.2	91.9	97.2	96.8	97.0	-	-	-	95.7	95.5	95.6
AnE_{LID}	94.16	98.46	96.26	99.41	97.69	98.54	-	-	-	97.97	97.90	97.91
Vietnamese–English		*(7219)*			*(16974)*			*(614)*			*(24807)*	
AnE	90.64	95.07	92.80	98.17	95.82	96.98	60.72	65.96	63.23	95.05	94.86	94.93

Table 6: Zero-shot LID results. SoTA results from Barik et al. (2019); Yirmibeşoğlu and Eryiğit (2018). No existing Vietnamese–English SoTA.

training sets of the LID task we are evaluating here, and collapsed NER training splits. It is possible that introducing these NER datasets brings a conflict of annotation guidelines. Alternatively it is possible that introducing additional data here was simply not necessary. Either way, we have shown here that AnE_{NER} is an optional NER module in our system that performs on-par with baseline.

We now proceed to evaluate performance on low-resource language pairs, for which AnE is not explicitly fine-tuned on any code-switching data. Table 6 gives results in P, R and F_1 measure on the Indonesian–English, Turkish–English and Vietnamese–English language pairs.

Zero-shot AnE outperforms SoTA classifiers fine-tuned directly (in the cross-validation setup) for the language pairs. F_1 measures in all categories are improvements over SoTA. For Indonesian–English code-switching, AnE is evaluated at overall $F_1 = 93.45\%$, outperforming the previous SoTA of 88.86%. The same holds for Turkish–English, where AnE_{LID} is evaluated at $F_1 = 97.9\%$ compared to the previous SoTA of 95.6% (significant figures/digits reduced to match reported SoTA).

Overall F_1 for Vietnamese–English is 94.93, but this is severely affected by the low score in the 'Other' category of $F_1 = 63.23\%$. This is because their 'X' category (which we collapse to 'Other') represents language-neutral words, rather than named entities/punctuation/emojis as our 'Other' here is targeted at. There is a mismatch here. Such labelled words only arise as a result of human intervention in their semi-automatic language identification process, which may be a factor. Another factor is that the Vietnamese–English data is the only corpus originating from recordings. Many of these ambiguous words arise from the conversational discourse setting not present in social media, like interjections and fillers. AnE is not able to effectively handle such words. We have still set baseline performance for Vietnamese–English code-switching identification.

We hypothesize the good zero-shot performance of AnE may be attributed to two factors. The first is the multilingual pre-training of XLM-RoBERTa. Monolingual training data in Indonesian, Turkish and Vietnamese as well as English is included in the multilingual pre-training data of XLM-RoBERTa. This may contribute to the performance of AnE in distinguishing these languages from English. The second factor is the way we formulated this task: distinguishing English from not English. This task formulation aimed to be independent of the other language.

This zero-shot evaluation shows that AnE performs well at identifying code-switched English amongst words of other languages not seen in the fine-tuning data. We consider this a good result, and it means AnE can be a baseline system for future research on code-switching between any language and English. It can also be a tool to quickly gather more data for low resource language pairs in code-switching research.

5.3 Correlation With Typological Similarity

We finish by investigating the connection between language typology and difficulty in recognizing code-switching. To this end, we used lexical similarity as a measure of linguistic similarity. We note that our language identification task is mostly lexical, in identifying the language of individual words. But for lexically similar languages it cannot be solved perfectly even with an ideal lexicon; interlingual homographs, words with the same surface form in two languages but different meanings, are one example for why.

A challenge in this investigation is what measure of difficulty in recognizing code-switching to use. We found earlier that overall F_1 score is heavily affected by how named entities and other words are annotated. Meanwhile, annotation schemes also

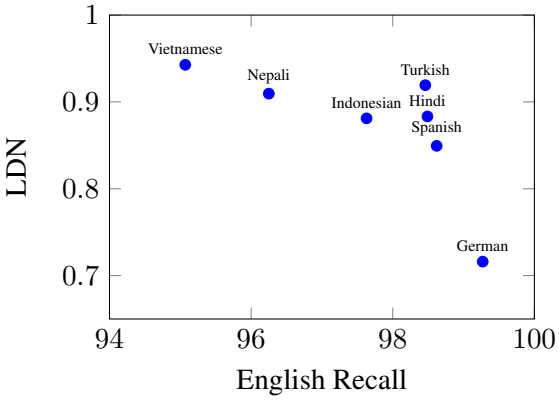


Figure 1: Normalized Levenshtein distance (LDN) between the ASJP wordlists in each of the languages, against English recall of AnE.

lead to some large variations between precision and recall. In the end, we decided to use the recall of English words as our measure. Ultimately, we anticipate this is the measure of most interest to those searching for code-switching in large corpora.

Figure 1 presents a plot of this recall measure against a measure of lexical similarity for each of the language pairs we explored in this work. Lexical similarity is computed as a distance using the ASJP corpus (Wichmann et al., 2022) and associated methods of Müller et al. (2010), but we acknowledge that no measure will be perfect.[7] We find a strong (ρ=-0.82) and significant (p=0.02) Spearman's correlation between lexical distance and English recall. But the correlation is negative, indicating that lexically similar languages are easier to distinguish. This is a surprising result, as one would expect that lexically dissimilar languages could be distinguished near-perfectly with a hashtable. We posit that an explanation may reside in the monolingual pre-training of the language model. It is plausible that the model learns representations that better distinguish lexically similar languages. An alternate hypothesis is that this correlation arises from the volume of pre-training data in each of the languages. It is also possible that the correlation is just a facet of the difficulty of each code-switching corpus we investigate.

6 Conclusion

In this work, we have presented a system (AnE) that distinguishes English words and words of other languages in multilingual text. On high-resource language pairs, the system underperforms language-pair-specific SoTA by a numerically small margin (always less than 1% absolute F_1). Meanwhile, it outperforms SoTA on low-resource language pairs, even though it was not trained on any code-switching of these language pairs. Analysis of our results revealed a negative correlation between lexical similarity and difficulty in recognizing code-switching, a surprising result which we leave to future work for further exploration. We believe our work bridges some of the resource-gap in code-switching research. We make it possible to compile new large-scale code-switching corpora of currently underrepresented language pairs. AnE is also a new and competitive baseline in code-switching identification research between any language and English.

Limitations

The main limitation of this work is in the language pairs AnE is able to support. The main motivation for this work was to make the most of existing high-resource code-switching data to support research on lower-resource language pairs in code-switching. We achieved this, but only for language pairs where one language is English.

There are of course many code-switching language pairs that do not involve English. But we found the data is not available today to train an AnE-type system to support those lines of research. For example, we would have wished to train a system that distinguished between code-switching of different language families, e.g. *Romance* vs. *notRomance*.

Ethics Statement

The Turkish–English and Vietnamese–English corpora we used were made available to us on our request to the authors of those works. The latter is a corpus of spoken code-switching, and hence comes with additional privacy constraints. We do not train any systems on that data, only using it for evaluation. All other corpora are publically available.

Acknowledgements

We are grateful to Andreas Vlachos for fruitful discussions and thank our four reviewers for their comments.

[7] We use the numbers released here.

References

Gustavo Aguilar, Fahad AlGhamdi, Victor Soto, Mona Diab, Julia Hirschberg, and Thamar Solorio. 2018a. 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*, pages 138–147, Melbourne, Australia. Association for Computational Linguistics.

Gustavo Aguilar, Fahad AlGhamdi, Victor Soto, Thamar Solorio, Mona Diab, and Julia Hirschberg, editors. 2018b. *Proceedings of the Third Workshop on Computational Approaches to Linguistic Code-Switching*. Association for Computational Linguistics, Melbourne, Australia.

Gustavo Aguilar, Sudipta Kar, and Thamar Solorio. 2020. LinCE: A Centralized Benchmark for Linguistic Code-switching Evaluation. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 1803–1813, Marseille, France. European Language Resources Association.

Anab Maulana Barik, Rahmad Mahendra, and Mirna Adriani. 2019. Normalization of Indonesian-English code-mixed Twitter data. In *Proceedings of the 5th Workshop on Noisy User-generated Text (W-NUT 2019)*, pages 417–424, Hong Kong, China. Association for Computational Linguistics.

Mirjam Broersma. 2009. Triggered codeswitching between cognate languages. *Bilingualism: Language and Cognition*, 12(4):447–462.

Mirjam Broersma and Kees De Bot. 2006. Triggered codeswitching: A corpus-based evaluation of the original triggering hypothesis and a new alternative. *Bilingualism: Language and cognition*, 9(1):1–13.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Ching-Ting Chang, Shun-Po Chuang, and Hung-Yi Lee. 2019. Code-Switching Sentence Generation by Generative Adversarial Networks and its Application to Data Augmentation. In *Proc. Interspeech 2019*, pages 554–558.

Michael G Clyne. 1980. Triggering and language processing. *Canadian Journal of Psychology/Revue canadienne de psychologie*, 34(4):400.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Margaret Deuchar. 2009. The miami corpus: Documentation file.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Penelope Gardner-Chloros. 2020. Contact and code-switching. *The handbook of language contact*, pages 181–199.

Deepak Gupta, Asif Ekbal, and Pushpak Bhattacharyya. 2020. A semi-supervised approach to generate the code-mixed text using pre-trained encoder and transfer learning. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2267–2280.

Asha Hegde, F Balouchzahi, Sharal Coelho, Shashirekha H L, Hamada A Nayel, and Sabur Butt. 2024. Coli@fire2023: Findings of word-level language identification in code-mixed tulu text. In *Proceedings of the 15th Annual Meeting of the Forum for Information Retrieval Evaluation*, FIRE '23, page 25–26, New York, NY, USA. Association for Computing Machinery.

Ahmad Fathan Hidayatullah, Atika Qazi, Daphne Teck Ching Lai, and Rosyzie Anna Apong. 2022. A systematic review on language identification of code-mixed text: Techniques, data availability, challenges, and framework development. *IEEE Access*, 10:122812–122831.

Navya Jose, Bharathi Raja Chakravarthi, Shardul Suryawanshi, Elizabeth Sherly, and John P. McCrae. 2020. A survey of current datasets for code-switching research. In *2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS)*, pages 136–141.

Diederik P. Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980.

Dau-Cheng Lyu and Ren-Yuan Lyu. 2008. Language identification on code-switching utterances using multiple cues. In *Ninth Annual Conference of the International Speech Communication Association*.

Manuel Mager, Özlem Çetinoğlu, and Katharina Kann. 2019. Subword-level language identification for intra-word code-switching. *arXiv preprint arXiv:1904.01989*.

Deepthi Mave, Suraj Maharjan, and Thamar Solorio. 2018. Language Identification and Analysis of Code-Switched Social Media Text. In *Proceedings of the Third Workshop on Computational Approaches to Linguistic Code-Switching*, pages 51–61, Melbourne, Australia. Association for Computational Linguistics.

Giovanni Molina, Fahad AlGhamdi, Mahmoud Ghoneim, Abdelati Hawwari, Nicolas Rey-Villamizar, Mona Diab, and Thamar Solorio. 2016. Overview for the second shared task on language identification in code-switched data. In *Proceedings of the Second Workshop on Computational Approaches to Code Switching*, pages 40–49, Austin, Texas. Association for Computational Linguistics.

André Müller, Søren Wichmann, Viveka Velupillai, Cecil H Brown, Pamela Brown, Sebastian Sauppe, Eric W Holman, Dik Bakker, Johann-Mattis List, Dmitri Egorov, et al. 2010. Asjp world language tree of lexical similarity: Version 3 (july 2010). *Retrieved*, 10(19):2015.

Carol Myers-Scotton. 1992. Comparing codeswitching and borrowing. *Journal of Multilingual and Multicultural Development*, 13(1-2):19–39.

Ravindra Nayak and Raviraj Joshi. 2022. L3Cube-HingCorpus and HingBERT: A code mixed Hindi-English dataset and BERT language models. In *Proceedings of the WILDRE-6 Workshop within the 13th Language Resources and Evaluation Conference*, pages 7–12, Marseille, France. European Language Resources Association.

Li Nguyen and Christopher Bryant. 2020. CanVEC - the canberra Vietnamese-English code-switching natural speech corpus. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4121–4129, Marseille, France. European Language Resources Association.

Li Nguyen, Christopher Bryant, Sana Kidwai, and Theresa Biberauer. 2021. Automatic language identification in code-switched hindi-english social media text. *Journal of Open Humanities Data*, 7.

Doreen Osmelak and Shuly Wintner. 2023. The denglisch corpus of German-English code-switching. In *Proceedings of the 5th Workshop on Research in Computational Linguistic Typology and Multilingual NLP*, pages 42–51, Dubrovnik, Croatia. Association for Computational Linguistics.

Lance Ramshaw and Mitch Marcus. 1995. Text chunking using transformation-based learning. In *Third Workshop on Very Large Corpora*.

Mohd Sanad Zaki Rizvi, Anirudh Srinivasan, Tanuja Ganu, Monojit Choudhury, and Sunayana Sitaram. 2021. GCM: A toolkit for generating synthetic code-mixed text. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 205–211, Online. Association for Computational Linguistics.

Caroline Sabty, Islam Mesabah, Özlem Çetinoğlu, and Slim Abdennadher. 2021. Language identification of intra-word code-switching for arabic–english. *Array*, 12:100104.

Safaa Shehadi and Shuly Wintner. 2022a. Identifying code-switching in arabizi. In *Proceedings of the The Seventh Arabic Natural Language Processing Workshop (WANLP)*, pages 194–204.

Safaa Shehadi and Shuly Wintner. 2022b. Identifying code-switching in Arabizi. In *Proceedings of the The Seventh Arabic Natural Language Processing Workshop (WANLP)*, pages 194–204, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Kushagra Singh, Indira Sen, and Ponnurangam Kumaraguru. 2018. Language identification and named entity recognition in Hinglish code mixed tweets. In *Proceedings of ACL 2018, Student Research Workshop*, pages 52–58, Melbourne, Australia. Association for Computational Linguistics.

Thamar Solorio, Elizabeth Blair, Suraj Maharjan, Steven Bethard, Mona Diab, Mahmoud Ghoneim, Abdelati Hawwari, Fahad AlGhamdi, Julia Hirschberg, Alison Chang, and Pascale Fung. 2014. Overview for the first shared task on language identification in code-switched data. In *Proceedings of the First Workshop on Computational Approaches to Code Switching*, pages 62–72, Doha, Qatar. Association for Computational Linguistics.

Thamar Solorio, Shuguang Chen, Alan W. Black, Mona Diab, Sunayana Sitaram, Victor Soto, Emre Yilmaz, and Anirudh Srinivasan, editors. 2021. *Proceedings of the Fifth Workshop on Computational Approaches to Linguistic Code-Switching*. Association for Computational Linguistics, Online.

Thamar Solorio, Monojit Choudhury, Kalika Bali, Sunayana Sitaram, Amitava Das, and Mona Diab, editors. 2020. *Proceedings of the The 4th Workshop on Computational Approaches to Code Switching*. European Language Resources Association, Marseille, France.

Victor Soto, Nishmar Cestero, and Julia Hirschberg. 2018. The role of cognate words, pos tags and entrainment in code-switching. In *Interspeech*, pages 1938–1942.

Igor Sterner and Simone Teufel. 2023. TongueSwitcher: Fine-grained identification of German-English code-switching. In *Proceedings of the 6th Workshop on Computational Approaches to Linguistic Code-Switching*, pages 1–13, Singapore. Association for Computational Linguistics.

Søren Wichmann, Eric W. Holman, and Cecil H. Brown. 2022. The asjp database (version 20). http://asjp.clld.org/.

Genta Winata, Sudipta Kar, Marina Zhukova, Thamar Solorio, Mona Diab, Sunayana Sitaram, Monojit Choudhury, and Kalika Bali, editors. 2023. *Proceedings of the 6th Workshop on Computational Approaches to Linguistic Code-Switching*. Association for Computational Linguistics, Singapore.

Shuly Wintner, Safaa Shehadi, Yuli Zeira, Doreen Osmelak, and Yuval Nov. 2023. Shared Lexical Items as Triggers of Code Switching. *Transactions of the Association for Computational Linguistics*, 11:1471–1484.

Zeynep Yirmibeşoğlu and Gülşen Eryiğit. 2018. Detecting code-switching between turkish-english language pair. In *Proceedings of the 2018 EMNLP Workshop W-NUT: The 4th Workshop on Noisy User-generated Text*, pages 110–115.

Zheng Xin Yong, Ruochen Zhang, Jessica Forde, Skyler Wang, Arjun Subramonian, Holy Lovenia, Samuel Cahyawijaya, Genta Winata, Lintang Sutawika, Jan Christian Blaise Cruz, Yin Lin Tan, Long Phan, Long Phan, Rowena Garcia, Thamar Solorio, and Alham Aji. 2023. Prompting multilingual large language models to generate code-mixed texts: The case of south East Asian languages. In *Proceedings of the 6th Workshop on Computational Approaches to Linguistic Code-Switching*, pages 43–63, Singapore. Association for Computational Linguistics.

Ruochen Zhang, Samuel Cahyawijaya, Jan Christian Blaise Cruz, Genta Winata, and Alham Aji. 2023. Multilingual large language models are not (yet) code-switchers. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12567–12582, Singapore. Association for Computational Linguistics.

Yuan Zhang, Jason Riesa, Daniel Gillick, Anton Bakalov, Jason Baldridge, and David Weiss. 2018. A fast, compact, accurate model for language identification of codemixed text. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 328–337, Brussels, Belgium. Association for Computational Linguistics.