Detecting Code-Switching between Turkish-English Language Pair

Zeynep Yirmibeşoğlu
Dep. of Computer Engineering
Istanbul Technical University
Maslak, Istanbul 34369
yirmibesogluz@itu.edu.tr

Gülşen Eryiğit
Dep. of Computer Engineering
Istanbul Technical University
Maslak, Istanbul 34369
gulsen.cebiroglu@itu.edu.tr

Abstract

Code-switching (usage of different languages within a single conversation context in an alternative manner) is a highly increasing phenomenon in social media and colloquial usage which poses different challenges for natural language processing. This paper introduces the first study for the detection of Turkish-English code-switching and also a small test data collected from social media in order to smooth the way for further studies. The proposed system using character level n-grams and conditional random fields (CRFs) obtains 95.6% micro-averaged F1-score on the introduced test data set.

1 Introduction

Code-switching is a common linguistic phenomenon generally attributed to bilingual communities but also highly observed among white collar employees. It is also treated as related to higher education in some regions of the world (e.g. due to foreign language usage at higher education). Although the social motivation of code-switching usage has been still under investigation and there exist different reactions to it (Hughes et al., 2006; Myers-Scotton, 1995), the challenges caused by its increasing usage in social media are not negligible for natural language processing studies focusing on this domain.

Social media usage has increased tremendously, bringing with it several problems. Analysis and information retrieval from social media sources are difficult, due to usage of a noncanonical language (Han and Baldwin, 2011; Melero et al., 2015). The noisy character of social media texts often require text normalization, in order to prepare social media texts for data analysis. Eryiğit and Torunoğlu-Selamet (2017) is the first study which introduces a social media text normalization approach for Turkish. In this study, similar to Han and Baldwin (2011) their candidate word (solution) generation stage comes after an initial ill-formed word detection stage where they use a Turkish morphological analyzer as the language validator. Although this approach works quite well for Turkish posts, it is obvious that it would encounter difficulties in case of code-switching where the language validator would detect every foreign word as ill-formed and the normalizer would try to propose a candidate correction for each of these. A similar situation may be observed at the behavior of spelling checkers within text editors. These also detect the foreign words (purposely written) as out of vocabulary and insist on proposing a candidate correction which makes the detection of actual spelling errors difficult for the users.

In recent years, the use of code-switching between Turkish and English has also become very frequent specifically in daily life conversations and social media posts of white collars and youth population. Ex. (1) introduces such an example which is not unusual to see.

(1) ≪Original code-switched version≫

Serverlarımızın **update** işlemleri için bu **domain**deki **expert** arayışımız devam etmektedir.

≪Turkish version and literal translation≫

Sunucularımızın (*of our servers*) güncelleme (*update*) işlemleri (*process*) için (*for*), bu (*this*) alandaki (*on domain*) uzman (*expert*) arayışımız (*search*) devam etmektedir (*continues*).

≪English version≫

For the update processes of our servers, we keep on searching an expert on this domain.

To the best of our knowledge, this is the first study working on automatic detection of code-switching between Turkish and English. We introduce a small test data set composed of 391 social media posts each consisting of code-switched

sentences and their word-by-word manual annotation stating either the word is Turkish or English. The paper presents our first results on this data set which is quite promising with a 95.6% micro average F1-score. Our proposed system uses conditional random fields using character n-grams and word look-up features from monolingual corpora.

2 Related Work

Code-switching is a spoken and written phenomenon. Hence, its investigation by linguists had started long before the Internet era, dating to 1950s (Solorio et al., 2014). However, code-switching researches concerning Natural Language Processing has started more recently, with the work of Joshi (1982), where a "formal model for intra-sentential code-switching" is introduced.

Analysis of code-switched data requires an annotated, multilingual corpus. Although collection of code-switched social media data is not an easy task, there has been worthy contributions. A Turkish-Dutch corpus (Nguyen and Doğruöz, 2013), a Bengali-English-Hindi corpus (Barman et al., 2014), Modern Standard Arabic - Dialectal Arabic, Mandarin - English, Nepali-English, and Spanish-English corpora for the First and Second Workshops on Computational Approaches to Code-switching (Solorio et al., 2014; Molina et al., 2016), a Turkish-German corpus (Özlem Çetinoğlu, 2016), a Swahili-English corpus (Piergallini et al., 2016) and an Arabic-Moroccan Darija corpus (Samih and Maier, 2016) were introduced. Social media sources are preferred, due to the fact that social media users are not aware that their data are being analyzed and thus generate text in a more natural manner (Çetinoğlu et al., 2016). To our knowledge, a Turkish-English code-switching social media corpus has not yet been introduced.

Word-level language identification of code-switched data has proved to be a popular research area with the ascent of social media. Das and Gambäck (2013) applied language detection to Facebook messages in mixed English-Bengali and English-Hindi. Chittaranjan et al. (2014) carried out the task of language detection for code-switching feeding character n-grams to CRF, with addition to lexical, contextual and other special character features, and reached 95% labeling accuracy. Nguyen and Doğruöz (2013) identified Dutch-Turkish words using character n-grams and dictionary lookup as CRF features along with contextual features, reaching 98% accuracy, and introducing new methods to measure corpus complexity. These researches mostly depend on monolingual training data (Solorio et al., 2014). As opposed to monolingual training data, Lignos and Marcus (2013) used code-switched data for language modeling where they use a Spanish data set containing 11% of English code-switched data. However, the usage of code-switched data for training is problematic, since its size is generally low, and may be insufficient for training (Maharjan et al., 2015).

Shared tasks on code-switching (Solorio et al., 2014; Molina et al., 2016) contributed greatly to the research area. First Workshop on Computational Approaches to Code-switching (FWCAC) showed that, when typological similarities are high between the two languages (Modern Standard Arabic-Dialectal Arabic (MSA-DA) for instance), and they share a big amount of lexical items, language identification task becomes considerably difficult (Solorio et al., 2014). It is easier to define languages when the two are not closely related (Nepali-English for instance).

3 Language Identification Models

This section presents our word-level identification models tested on Turkish-English language pair.

3.1 Character N-gram Language Modeling

Our first model "**Ch.n-gram**" uses SRI Language Modeling Toolkit (SRILM) for character n-gram modeling, with Witten-Bell smoothing (Stolcke, 2002). Unigrams, bigrams and trigrams (n=1,2,3) are extracted from Turkish and English training corpora (ETD, TNC and TTC to be introduced in §4).

In order to observe how corpora with different sources (formal or social media) affect language modeling and word-level language identification on social media texts, TNC and TTC are paired with the ETD, and the model perplexities are calculated against the code-switched corpus (CSC). Language labels are decided upon the comparison of English and Turkish model perplexities for each token in the test set.

3.2 Conditional Random Fields (CRF)

Conditional Random Fields (CRF) perform effectively in the sequence labeling problem for many

NLP tasks, such as Part-of-Speech (POS) tagging, information extraction and named entity recognition (Lafferty et al., 2001). CRF method was employed by Chittaranjan et al. (2014) for word-level language detection, using character n-gram probabilities among others as a CRF feature, reaching 80% - 95% accuracy in different language pairs. In this research we also experiment with CRF for word-level language identification, where language tagging is considered as a sequence labeling problem of labeling a word either with English or Turkish language tags.

Our first CRF model "**CRF**†" uses lexicon lookup (LEX), character n-gram language model (LM) features and the combination of these for the current and neighboring tokens (provided as feature templates to the used CRF tool (Kudo, 2005)). LEX features are two boolean features stating the presence of the current token in the English (ETD) or Turkish dictionary (TNC or TTC). LM feature is a single two-valued (T or E) feature stating the label assigned by our previous (*Ch.n-gram*) model introduced in §3.1.

Turkish is an agglutinative language. Turkish proper nouns are capitalized and an apostrophe is inserted between the noun and any following inflectional suffix. It is frequently observed that code-switching people apply the same approach while using foreign words in their writings. Ex. (2) provides such an example usage:

(2) **action**'lar ≪code-switched version≫
eylemler ≪Turkish version≫
actions ≪English version≫

In such circumstances, it is hard for our character-level and lexicon look-up models to assign a correct tag where an intra-word code-switching occurs and the apostrophe sign may be a good clue for detecting these kinds of usages. In order to reflect this know-how to our machine learning model, we added new features (APOS) to our last model "**CRF**$^{\phi}$" (in addition to previous ones). APOS features are as follows: a boolean feature stating whether the token contains an apostrophe (') sign or not, a feature stating the language tag (E or T) assigned by *ch.n-gram* model to the word sub-part appearing before the apostrophe sign (this feature is assigned an 'O' (other) tag if the previous boolean feature is 0) and a final feature which is similar to the previous one but this time stating whether this sub-part appears in one of the language dictionaries (E/T/O).

4 Data

Our character-level n-gram models were trained on monolingual Turkish and English corpora retrieved from different sources. We also collected and annotated a Turkish-English code-switched test data-set and used it both for testing of our n-gram models and training (via cross-validation) of our sequence labeling model.

The monolingual English training data (ETD) was acquired from the Leipzig Corpora Collection (Goldhahn et al., 2012), containing English text from news resources, incorporating a formal language, with 10M English tokens. For the Turkish training data, two different corpora were used. The first corpus was artificially created using the word frequency list of the Turkish National Corpus (TNC) Demo Version (Aksan et al., 2012). TNC mostly consists of formally written Turkish words. Second Turkish corpus (TTC) (6M tokens) was extracted using the Twitter API aiming to obtain a representation of the non-canonical user-generated context.

For the code-switched test corpus (CSC), 391 posts all of which containing Turkish-English code-switching were collected from Twitter posts and Ekşi Sözlük website[1]. The data was cross-annotated by two human annotators. A baseline assigning the label "Turkish" to all tokens in this dataset would obtain a 72.6% accuracy score and a 42.1% macro-average F1-measure. Corpus statistics and characteristics are provided in Table 1.[2]

5 Experiments and Discussions

We evaluate our token-level language identification models using precision, recall and F_1 measures calculated separately for both language classes (Turkish and English). We also provide micro[3] and macro averaged F_1 measures for each model.

Table 2 provides two baselines: the first row "base_T" (introduced in §4) provides the scores of

[1] an online Turkish forum, containing informal, user-generated information

[2] All punctuations (except for " ' " and "-"), smileys and numbers were removed from the corpora using regular expressions (regex) and all characters were lowercased.

[3] Since the accuracy scores in this classification scenario are most of the time the same (except for baseline model scores provided in §4) with micro-averaged F_1 measures, we do not provide them separately in Table 2.

	English tokens	Turkish tokens	Total tokens	Language Type	Use
TNC	-	10,943,259	10,943,259	Formal	Training of Turkish language model
TTC	-	5,940,290	5,940,290	Informal	Training of Turkish language model
ETD	10,799,547	-	10,799,547	Formal	Training of English language model
CSC	1488	3942	5430	Informal	Testing & Training of sequence models

Table 1: Corpus Characteristics

System	LM/ Dict.	Turkish			English			Avg. F_1	
		P	R	F_1	P	R	F_1	Micro	Macro
base_T	-	72.6%	100.0%	84.1%	0.0%	0.0%	0.0%	61.1%	42.1%
base_LL	ETD-TNC	91.4%	98.7%	94.9%	95.5%	75.5%	84.4%	92.0%	89.6%
Ch.n-gram	ETD-TNC	98.1%	88.4%	93.0%	75.6%	95.5%	84.4%	90.6%	88.7%
Ch.n-gram	ETD-TTC	95.9%	94.1%	95.0%	85.1%	89.4%	87.2%	92.9%	91.1%
CRF†	ETD-TNC	96.3%	97.2%	96.7%	91.9%	89.6%	90.6%	95.0%	93.7%
CRF†	ETD-TTC	96.3%	96.9%	96.6%	91.2%	90.3%	90.7%	95.0%	93.6%
CRF$^{\phi}$	ETD-TNC	97.2%	96.8%	**97.0%**	91.7%	92.2%	**91.9%**	**95.6%**	**94.5%**
CRF$^{\phi}$	ETD-TTC	96.8%	96.6%	96.7%	91.5%	90.9%	91.1%	95.1%	93.9%

Table 2: Token-level language identification results.
LM/Dict. refers to the data used as dictionaries and training data for n-gram language models.

the baseline model which assigns the label "Turkish" to all tokens and the second row provides the results of a rule-based lexicon lookup (base_LL) which assigns the language label for each word by searching it in TNC and ETD used as Turkish and English dictionaries. If a word occurs in both or none of these dictionaries, it is tagged as Turkish by default.

We observe from the results that the character-level n-gram models trained on a formal data set (TNC) fall behind our second baseline (with 88.7% macro avg. F_1) whereas the one trained on social media data (TTC) performs better (91.1%). It can also be observed that the performances of character n-gram language models turned out to be considerably high, aided by the fact that Turkish and English are morphologically distant languages and contain differing alphabetical characters such as "ş,ğ,ü,ö,ç,ı"in Turkish and "q,w,x" in English.

CRF models' performances are calculated via 10 fold cross-validation over code-switched corpus (CSC). One may see from the table that all of our CRF models perform higher than our baselines and character n-gram models. The best performances (95.6% micro and 94.5% macro avg. F_1) are obtained with **CRF$^{\phi}$** trained with LEX + LM + APOS features. Contrary to the above findings with character level n-gram models, we see that **CRF$^{\phi}$** performs better when TNC is used for character-level n-gram training and look-up. The use of TTC (monolingual Turkish data collected from social media) was revealing better results in Ch.n-gram models and similar results in CRF†. This may be attributed to the fact that our hypothesis regarding the use of apostrophes in code-switching of Turkish reveals a good point and the validation of the word sub-part before the apostrophe sign (from a formally written corpus - TNC) brings out a better modeling.

6 Conclusion

In this paper, we presented the first results on code-switching detection between Turkish-English languages. With the motivation of social media analysis, we introduced the first data set which consists 391 posts with 5430 tokens (having ∼30% English words) collected from social media posts[4]. Our first trials with conditional random fields revealed promising results with 95.6% micro-average and 94.5 macro-average F_1 scores. We see that there is still room for improvement for future studies in order to increase the relatively low F_1 (91.9%) scores on English. As future works, we aim to increase the corpus size

[4]The code-switched corpus is available via http://tools.nlp.itu.edu.tr/Datasets (Eryiğit, 2014)

and to test with different sequence models such as LSTMs.

References

Yeşim Aksan, Mustafa Aksan, Ahmet Koltuksuz, Taner Sezer, Ümit Mersinli, Umut Ufuk Demirhan, Hakan Yılmazer, Gülsüm Atasoy, Seda Öz, İpek Yıldız, and Özlem Kurtoğlu. 2012. Construction of the Turkish national corpus (tnc). In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC 2012)*, Istanbul, Turkey. European Language Resources Association (ELRA).

Utsab Barman, Amitava Das, Joachim Wagner, and Jennifer Foster. 2014. 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 (EMNLP 2014)*, pages 13–23. Association for Computational Linguistics.

Özlem Çetinoğlu, Sarah Schulz, and Ngoc Thang Vu. 2016. Challenges of computational processing of code-switching. In *Proceedings of the Second Workshop on Computational Approaches to Code Switching (EMNLP 2016)*, pages 1–11, Austin, Texas. Association for Computational Linguistics.

Özlem Çetinoğlu. 2016. A Turkish-German code-switching corpus. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France. European Language Resources Association (ELRA).

Gokul Chittaranjan, Yogarshi Vyas, Kalika Bali, and Monojit Choudhury. 2014. 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 (EMNLP 2014)*, pages 73–79. Association for Computational Linguistics.

Amitava Das and Björn Gambäck. 2013. Code-mixing in social media text: The last language identification frontier? *Traitement Automatique des Langues (TAL): Special Issue on Social Networks and NLP*, 54(3).

Gülşen Eryiğit and Dilara Torunoğlu-Selamet. 2017. Social media text normalization for Turkish. *Natural Language Engineering*, 23(6):835–875.

Gülşen Eryiğit. 2014. ITU Turkish NLP web service. In *Proceedings of the Demonstrations at the 14th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, Gothenburg, Sweden. Association for Computational Linguistics.

Dirk Goldhahn, Thomas Eckart, and Uwe Quasthoff. 2012. Building large monolingual dictionaries at the Leipzig corpora collection: From 100 to 200 languages. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC 2012)*.

Bo Han and Timothy Baldwin. 2011. Lexical normalisation of short text messages: Makn sens a# twitter. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pages 368–378. Association for Computational Linguistics.

Claire E Hughes, Elizabeth S Shaunessy, Alejandro R Brice, Mary Anne Ratliff, and Patricia Alvarez McHatton. 2006. Code switching among bilingual and limited English proficient students: Possible indicators of giftedness. *Journal for the Education of the Gifted*, 30(1):7–28.

Aravind K. Joshi. 1982. Processing of sentences with intra-sentential code-switching. In *Proceedings of the 9th Conference on Computational Linguistics - Volume 1*, COLING '82, pages 145–150, Czechoslovakia. Academia Praha.

Taku Kudo. 2005. Crf++: Yet another crf toolkit. *http://crfpp. sourceforge. net/*.

John Lafferty, Andrew McCallum, and Fernando Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the 18th International Conference on Machine Learning*, pages 282–289. Morgan Kaufmann, San Francisco, CA.

Constantine Lignos and Mitch Marcus. 2013. Toward web-scale analysis of codeswitching. In *87th Annual Meeting of the Linguistic Society of America*.

Suraj Maharjan, Elizabeth Blair, Steven Bethard, and Thamar Solorio. 2015. Developing language-tagged corpora for code-switching tweets. In *Proceedings of LAW IX - The 9th Linguistic Annotation Workshop*, pages 72–84, Denver, Colorado.

Maite Melero, Marta.R. Costa-Jussà, P. Lambert, and M. Quixal. 2015. Selection of correction candidates for the normalization of Spanish user-generated content. *Natural Language Engineering*, FirstView:1–27.

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 (EMNLP 2016)*, pages 40–49. Association for Computational Linguistics.

Carol Myers-Scotton. 1995. *Social motivations for codeswitching: Evidence from Africa*. Oxford University Press.

Dong Nguyen and A. Seza Doğruöz. 2013. Word level language identification in online multilingual communication. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 857–862, Seattle, Washington, USA. Association for Computational Linguistics.

Mario Piergallini, Rouzbeh Shirvani, Gauri S. Gautam, and Mohamed Chouikha. 2016. Word-level language identification and predicting codeswitching points in Swahili-English language data. In *Proceedings of the Second Workshop on Computational Approaches to Code Switching (EMNLP 2016)*, pages 21–29. Association for Computational Linguistics.

Younes Samih and Wolfgang Maier. 2016. An Arabic-Moroccan Darija code-switched corpus. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France. European Language Resources Association (ELRA).

Thamar Solorio, Elizabeth Blair, Suraj Maharjan, Steven Bethard, Mona T. 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 (EMNLP 2014)*, pages 62–72, Doha, Qatar.

Andreas Stolcke. 2002. Srilm – an extensible language modeling toolkit. In *Proceedings of the 7th International Conference on Spoken Language Processing (ICSLP 2002)*, pages 901–904.