International Journal of Applied Engineering Research ISSN 0973-4562 Volume 13, Number 12 (2018) pp. 10339-10342
© Research India Publications. http://www.ripublication.com

A Short Review on the Language Identification Systems of Social Media Post

Priyadarshini Lamabam

Post Graduate Computer Science Teacher, Kendriya Vidhyalaya No.1 Lamphelpat,
Nagamapal Kangjabi Leirak, Imphal West, Manipur-795004 India.

Abstract:

Automatic language identification is one of the most pre-requisite tasks of natural language processing. The various Natural language processing (NLP) tools are developed based on this system. Various Natural language processing tools have been developed for processing the formal texts but very few works have been reported when language is not the formal text but the informal text that arises from social media. This is due to the peculiar nature of the social media data which challenges the task of the NLP research community. This paper has given a brief review on the language identification systems developed for social media with special reference to twitter and facebook.

Keywords: Natural Language Processing, social media, code-mixing, language identification

INTRODUCTION

Social media are computer-mediated tools that help the people, companies and other organizations in exchanging information, career interests, ideas, stories and pictures/videos in virtual communities and networks. It includes a wide range of Web sites and Internet-based services that allow users to create content and interact with other users. It is found that much of the social media interaction is personal and expressed between two people or among a group. They used either first person (“I”) and second person(“You”.”we”).The way people communicate in social media directly contrasts with the news or brand posts, that are likely written with a more detached, omniscient tone. The type of texts posted in social media are informal in nature.

Language identification of the texts laid the foundation for the development of the various NLP tools. And many state-of-the art technologies have been developed by the researchers for processing the formal texts written in most popular languages. Apart from working on formal documents, the NLP researchers have extended their focus on the data obtained from the multi-party chats, discussion forums, blogs, and online reviews. But within the last decade, there has been a rapid expansion of their work to a new direction that cover a new social media content—microblogs eg.,Twitter, social networks eg., Facebook, comments on news articles, captions on user-contributed images such as on Flickr, and forums dedicated to specialized topics and needs (e.g., health and online education). Therefore, the need for language identification algorithms that could process the informal data available on social media has been given utmost importance. Unfortunately, it is found that they do not perform well on social media due to the presence of lexical borrowings, creative spellings and phonetic typing. The posts from facebook and twitter represent different characteristics that make the language identification task more challenging. They consists of hashtags, user mentions, links, emojis. The social media participants likely varies in their language when they post on facebook or twitter. They tend to shorten and/or encode many words in the form of chatspeak, while introducing typos and misspellings and it deviates their text from the standard spellings. Although English is still by far the most popular language in Social Media Content, its dominance is receding. And a new type of content has given birth in Social Media and that is called code-mixing or code-switching where the users communicate using more than one language. This new concept has brought new research challenges in developing NLP tools for the research community. Code-mixing or code-switching has been recognized as a byproduct of two or more languages since 1980s.

After the various investigations on language identification for half a century as reported in [1] and that of computational analysis of code switching for several decades according to [2], we find that only few works on automatic language identification for multilingual code-mixed texts have been reported. The researchers have performed the various investigations to find out why code-mixing occurs in social media. Linguistic motivations for sociological and conversational necessity in highly bilingual society influence the people for code-mixing [3].Inter-sentential, intra-sentential and intra-word code mixing were some of the types of code-mixing that were described according to the researchers. In [4], researchers showed that the people who used facebook tend to mainly use inter-sentential switching (59%) over intra-sentential (33%) and tag switching (8%), and 45% of the switching is triggered by real lexical needs, 40% for talking about a particular topic, and 5% is contributed by content clarification. Very few works have been noted on the language identification of social media text. After a comprehensive study on them, some works have been reported in this paper.

Section 2 discusses about the various language identification tools developed for twitter and facebook data. The paper is concluded in section 3 highlighting the future work.

RELATED WORK

Research in the language identification of short text has increased especially in recent years, with the advent of social media and microblogs. Tromp and Pechenizkiy [5] proposed a graph-based n-gram approach for tweet language identification. They have used datasets with monolingual tweets in six languages, which led to the achievement of performances between 95% and 98%. Their work is extended in [6] with the proposal of several linguistically-motivated modifications to their algorithm and thereby achieving 99.8% accuracy.

The researchers in [7] have addressed the specific problem of detecting the two types of the Portuguese language-European and Brazilian available from Twitter micro-blogging data. They followed an automatic classification approach that uses a naive Bayes Classifier with 95% accuracy.

In [8] they used Twitter dataset with tweets in five major European languages: Dutch, English, French, German, and Spanish. They gave their focus specifically on language identification from Twitter messages by augmenting the standard methods with LangID priors based on the previous messages of the user and the content of links used in messages.

In [9] the researchers have made experiments on film subtitles in22 languages using a Bayesian classifier that properly identifies very short texts. The researchers in [10] examine LangID for creating language specific twitter collections from twitter datasets with tweets consisting of 9 languages that uses Cyrillic, Arabic, and Devanagari scripts. They tested two language identification systems, viz., Logistic regression classifier (LogR) and Prediction by partial matching (PPM) by using textual features such as n-grams, and user metadata from Twitter, as well as Wikipedia as an external resource. They showed that by combining n-grams and user metadata, 98% accuracy could be achieved by their system in each subtask that deal with three languages.

In [11] they used wikipedia for training to classify tweets in five different languages. They have tested statistical language identifiers, based on character frequencies. A boot-strapping technique has been introduced that significantly improves the accuracy of the language identifier. For the language identification of the Indian languages like Hindi, Bengali, Marathi, Punjabi, Oriya, Telugu, Tamil, Malayalam and Kannada, they have dealt with the short texts in [12].Several language identification systems have been evaluated and applied to tweets in [13]. They have represented two datasets constructed by a conventional manual annotation approach and a novel mostly automated method. Three previously published datasets (CARTER, BERGSMA and TROMP) were used along with them. 8 off-the-shelf Language identification systems were compared and showed that a simple voting over three specific systems consistently outperforms with an F-score of 0.935.

In [14], automatic language identification with the dataset of Bengali-Hindi-English Facebook comments was presented and they have used systems such as dictionaries, Support Vector Machine (SVM) and Conditional Random Field (CRF). Owing to lack of transliterated dictionary for Bengali and Hindi and phonetically-typed nature, the training set words are used as dictionaries. British National Corpus (BNC), LexNormList and SemEvalTwitter along with training set words are used for English language. Experiments were performed on SVM and CRF with char-n grams, presence in dictionaries, length of words and capitalization as features.

The researchers in [15] showed the different techniques for the identification of English-Hindi and English-Bengali language mixed in Facebook posts using character n-grams, dictionaries and Support Vector Machine classifiers. For n-gram modeling, experiments were carried out on the training data for n = (1, 2, 3, 4, 5, 6, 7).A lexical normalization dictionary prepared for Twitter was used for English and for Bengali, the Samsad English-Bengali dictionary as reported in [16] and the Bengali lexicon transliterated into Romanized text using the Modified Joint-Source- Channel mode [17] were used.

N-gram with weights, dictionary-based, Minimum Edit Distance(MED)-based weight and word context information are being included as features in SVM. Taking the different datasets of Nepali-English and Spanish-English some word-level classification experiments was performed using dictionary-based method, linear kernel Support Vector Machines (SVMs) and a k-nearest neighbor approach in [18].The British National Corpus, lexical normalization dictionary and the training set words are used as dictionaries. For SVM system, they have used char n-grams, dictionary-based labels, length of words, capitalization and contextual clues as features.

A CRF-based system for language identification for four language pairs namely, English-Spanish, English-Nepali, English-Mandarin and Standard Arabic-Arabic Dialects have been developed in [19]. Their method uses lexical, contextual, character n-gram and special character features, that helps in the replication across languages.

In [20], another research has been reported on language identification using english-hindi code mixing from facebook. The corpus was created from the facebook pages of some popular public figures and also from BBC news corpus. They annotated the data using matrix, normalization, word origin, Named entities and POS tagging. Their analysis has shown a significant amount of CodeMixing of English in Hindi matrix and Hindi in English matrix. Hindi words embed in English using formulaic patterns of Nouns and Particles while English language get mixed with Hindi at various forms ranging from single words to multi-word phrases. They have also indicated that code-mixing in social media needs a deeper analysis of structural and discourse linguistics.

The researchers in [21] have done testing on the limits of the existing word level language identification systems such as linguini[22], polyglot[23], langid.py[24] and Compact Language Detector2 (CLD2)[25] after the preparation of synthetic code-mixed dataset of 28 languages. Due to lesser accuracy of the previous systems, they extended the existing algorithms to Random, MaxWeighted, CoverSet and Optimal. Random algorithm assigns a randomly chosen label to a word from a possible set of labels which are obtained by setting a threshold value on the confidence scores of the classifiers. Maxweighted assigns the label of the classifier with the highest confidence. CoverSet assumes that code-mixing happens only with a few language though there is no restriction in the number of languages. optimal algorithm compute the set of possible labels based on a threshold value. If the actual(gold standard) label of a word belongs to the set, then it is assigned as the tag for the word. The extended algorithms outperformed better than the existing algorithms significantly.

CONCLUSION

Social media has brought a new dimension in the field of natural language processing. Because the systems designed for the formal texts fail to perform well in the informal texts. And therefore the researchers have been trying to develop better systems that could process the informal texts with high accuracy. The various research work on the language identification of social media posts has been thoroughly presented in this paper. With the advent of code-mixing in social media the research has become more challenging due to the complex nature of the corpus involved. Moreover, due to lack of sufficient data not much work has been explored on the code-mixing that involves regional languages. Therefore collecting more amount of data for the various languages along with the development of various NLP tools for processing them will bring up the languages to the global platform.

REFERENCES

[1] E. Mark Gold, Language identification in the limit, Information and Control, 10(5):447–474,1967.

[2] Aravind K. Joshi, Processing of sentences with intra-sentential code-switching, in proceedings of the 9th International conference on Computational linguistics, pages 145–150, Prague, Czechoslovakia, July 1982, ACL.

[3] Peter Auer, Code-Switching in Conversation: Language, Interaction and Identity, Routledge,2013.

[4] Taofik Hidayat, An analysis of code switching used by facebookers (a case study in a social network site),BA Thesis, English Education Study Program, College of Teaching and Education (STKIP), Bandung, Indonesia, October 2012 .

[5] Tromp, E., Pechenizkiy, M.: Graph-based n-gram language identification on short texts. In: Proc. 20th Machine Learning conference of Belgium and The Netherlands, pp. 27–34 (2011)

[6] Vogel, J., Tresner-Kirsch, D.: Robust Language Identification in Short, Noisy Texts : Improvements to LIGA. In: Proceedings of the 3rd International Workshop on Mining Ubiquitous and Social Environments (MUSE), pp. 1–9. Bristol, UK (2012)

[7] Laboreiro, G., Bosnjak, M., Sarmento, L., Rodrigues, E.M., Oliveira, E.: Determining language variant in microblog messages. In: Proceedings of the 28th ACM/SIGAPP Symposium On Applied Computing, pp. 902–907. ACM (2013)

[8] Simon Carter, Wouter Weerkamp, and Manos Tsagkias.2013. Microblog language identification: Overcoming the limitations of short, unedited and idiomatic text. Language Resources and Evaluation, pages 1–21.

[9] Winkelmolen, F., Mascardi, V.: Statistical Language Identification of Short Texts. In: Proceedings of the 3rd International Conference on Agents and Artificial Intelligence, pp. 498–503. Rome, Italy (2011)

[10] Bergsma, S., McNamee, P., Bagdouri, M., Fink, C., Wilson, T.: Language identification for creating language-specific twitter collections. In: Workshop on Language in Social Media, pp. 65–74. ACL (2012)

[11] Moises Goldszmidt, Marc Najork, and Stelios Paparizos. 2013. Boot-strapping language identifiers for short colloquial postings. In Proceedings of the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECMLPKDD 2013), Prague, Czech Republic.

[12] Murthy, K.N., Kumar, G.B.: Language identification from small text samples. Journal of Quantitative Linguistics 13(1), 57–80 (2006)

[13] Lui, M., Baldwin, T.: Accurate language identification of twitter messages. In: Proceedings of the 5th Workshop on Language Analysis for Social Media (LASM), pp. 17–25. Association for Computational Linguistics, Gothenburg, Sweden (2014)

[14] Utsab Burman, Amitava Das, Joachim Wagner and Jennifer Foster, Code Mixing: A Challenge for language identification in the language of social media, in proceedings of the 2014 EMNLP, pages 13–23, Doha, Qatar, October. ACL. 1st Workshop on Computational Approaches to Code Switching

[15] Amitava Das and Bjorn Gamback, Identifying languages at the word level in code-mixed Indian social media text, in proceedings of the 11th International Conference on Natural Language Processing, pp.169-178, Goa, India, December 2014. Association for Computational Linguistics.

[16] Sailendra Bisvas, Samsad Bengali-English dictionary",Sahitya Samsad,Calcutta India,3 edition,2000.

[17] Amitava Das, Tanik Saikh, Tapabrata Mondal, Asif Ekbal and Sivaji Bandyopadhyay, "English to Indian languages machine transliteration system at NEWS 2010", in proceedings of the 48th ACL, pages 71–75, Uppsala, Sweden, July. ACL. 2nd Named Entities Workshop.

[18] Utsab Barman, Joachim Wagner, Grzegorz Chrupała and Jennifer Foster, "DCU-UVT: Word-Level Language Classification with Code-Mixed Data", in proceedings of The First Workshop on Computational

Approaches to Code Switching, EMNLP 2014, Conference on Empirical Methods in Natural Language Processing, Doha, Qatar. Association for Computational Linguistics.

[19] Gokul Chittaranjan, Yogarshi Vyas, Kalika Bali and Monojit 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,pages 73–79, October 25, 2014, Doha, Qatar. ACL.

[20] Kalika Bali,Jatin Sharma and Monojit Choudhury. 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, pages 116-126,October 25, 2014, Doha,Qatar.Association for Computational Linguistics.

[21] Spandana Gella, Kalika Bali and Monojit Choudhury. ye word kis lang ka hai bhai?Testing the Limits of Word level Language Identification In proceedings of the 11th International Conference on Natural Language Processing, pp.169-178,Goa, India, December 2014. Association for Computational Linguistics.

[22] John M Prager. Linguini: Language identification for multilingual documents in proceedings of the 32nd Annual Hawaii International Conference, pages 11pp. IEEE. In Systems Sciences, 1999. HICSS-32

[23] Marco Lui and Timothy Baldwin. Cross-domain feature selection for language identification in proceedings of 5th International Joint Conference on Natural Language Process ing. Citeseer. 2011

[24] langid.py. In https://github.com/saffsd/langid.py

[25] Compact Language Detector2. In https://code.google.com/p/cld2/