Overview for the First Shared Task on Language Identification in Code-Switched Data

Thamar Solorio
Dept. of Computer Science
University of Houston
Houston, TX, 77004
solorio@cs.uh.edu

Elizabeth Blair, Suraj Maharjan, Steven Bethard
Dept. of Computer and Information Sciences
University of Alabama at Birmingham
Birmingham, AL, 35294
{eablair,suraj,bethard}@uab.edu

Mona Diab, Mahmoud Gohneim, Abdelati Hawwari, Fahad AlGhamdi
Dept. of Computer Science
George Washington University
Washington, DC 20052
{mtdiab,mghoneim,abhawwari,fghamdi}@gwu.edu

Julia Hirschberg and **Alison Chang**
Dept. of Computer Science
Columbia University
New York, NY 10027
julia@cs.columbia.edu
ayc2135@columbia.edu

Pascale Fung
Dept. of Electronic & Computer Engineering
Hong Kong University of Science and Technology
Clear Water Bay, Kowloon, Hong Kong
pascale@ece.ust.hk

Abstract

We present an overview of the first shared task on language identification on code-switched data. The shared task included code-switched data from four language pairs: Modern Standard Arabic-Dialectal Arabic (MSA-DA), Mandarin-English (MAN-EN), Nepali-English (NEP-EN), and Spanish-English (SPA-EN). A total of seven teams participated in the task and submitted 42 system runs. The evaluation showed that language identification at the token level is more difficult when the languages present are closely related, as in the case of MSA-DA, where the prediction performance was the lowest among all language pairs. In contrast, the language pairs with the higest F-measure where SPA-EN and NEP-EN. The task made evident that language identification in code-switched data is still far from solved and warrants further research.

1 Introduction

The main goal of this language identification shared task is to increase awareness of the outstanding challenges in the automated processing of Code-Switched (CS) data and motivate more research in this direction. We define CS broadly as a communication act, whether spoken or written, where two or more languages are being used interchangeably. In its spoken form, CS has probably been around ever since different languages first came in contact. Linguists have studied this phenomenon since the mid 1900s. In contrast, the Natural Language Processing (NLP) community has only recently started to pay attention to CS, with the earliest work in this area dating back to Joshi's theoretical work proposing an approach to parsing CS data (Joshi, 1982) based on the Matrix and Embedded language framework. With the wide-spread use of social media, CS is now being used more and more in written language and thus we are seeing an increase in published papers dealing with CS. We are specifically interested in intrasentential code switched phenomena. As a result of this task, we have successfully created the first set of annotated data for several language pairs with a coherent set of labels across the languages. As the shared task results show, CS poses new research questions that warrant new NLP approaches, and thus we expect to see a significant increase in NLP work in the coming years addressing CS phenomena in data.

The shared task covers four language pairs and is focused on social media data. We provided participants with annotated data from Twitter for the

Proceedings of The First Workshop on Computational Approaches to Code Switching, pages 62–72,
October 25, 2014, Doha, Qatar.

language pairs: Modern Standard Arabic-Arabic dialects (MSA-DA), Mandarin-English (MAN-EN), NEP-EN (NEP-EN), and SPA-EN (SPA-EN). These language pairs represent a good variety in terms of language typology and relatedness among pairs. They also cover languages with different representation in terms of number of speakers world wide. Participants were asked to make predictions on unseen Twitter data for each language pair. We also provided participants with test data from a "surprise genre" with the objective of assessing the robustness of language identification systems to genre variation.

2 Task Description

The task consists of labeling each token/word in the input file with one of six labels: *lang1*, *lang2*, *other*, *ambiguous*, *mixed*, and named entities NE. The *lang1*, *lang2* labels refer to the two languages addressed in the subtask, for example for the language pair MSA-DA, *lang1* would be an MSA and *lang2* is DA. The *other* category is a label used to tag all punctuation marks, emoticons, numbers, and similar tokens that do not represent actual words in any of the given languages. The *ambiguous* label is for instances where it is not possible to assign a language with certainty, for example, a lexical form that belongs to both languages, appearing in a context that does not indicate one language over the other. The *mixed* category is for words composed of CS morphemes, such as the word snap*chateando* 'to chat' from SPA-EN, the word over*ai* from NEP-EN, or the word *hayqwlwn*[1] 'they will say', from MSA-DA, where the 'ha' is a DA future morpheme and the stem 'yqwlwn' is MSA.The *NE* label is included in this task in an effort to allow for a more focused analysis of CS data with the exclusion of proper nouns. NEs have a very different behavior than most other words in a language vocabulary and thus from our perspective they need to be identified to be handled properly.

Table 1 shows Twitter examples taken from the training data. The annotation guidelines are posted on the workshop website[2]. We post the ones used for SPA-EN as for the other language pairs the only differences are the examples provided.

[1]We use Buckwalter transliteration scheme `http://www.qamus.org/transliteration.htm`

[2]`http://emnlp2014.org/workshops/CodeSwitch/call.html`

Language Pair	Example
MSA-DA	*AlnhArdp AlsAEp 11 hAkwn Dyf >. HAfZ AlmyrAzy ElY qnAp drym* llHdyv En >wlwyAt Alvwrp fy AlmrHlp Al-HAlyp wqDyp tSHyH msAr Alvwrp Al<ElAmy (Today O'Clock 11 I_will_be [a_]guest[_of] Mr. Hafez AlMirazi on Channel Dream to_talk about [the_]priorities[_of] the_revolution in the_stage the_current and_[the_]issue[_of] correcting [the_]path[_of] the_revolution Media)
NEP-EN	My car at the workshop for a much needed repairs... *ABA pocket khali hune bho* (My car at the workshop for a much needed repairs… now my pocket will be empty)
SPA-EN	*Por primera vez veo a* @username actually being hateful! it was beautiful:) (For the first time I get to see @username actually being hateful! it was beautiful:)

Table 1: Examples of Twitter data used in the shared task.

3 Related Work

In the past, most language identification research has been done at the document level. Some researchers, however, have developed methods to identify languages within multilingual documents (Singh and Gorla, 2007; Nguyen and Doğruöz, 2013; King and Abney, 2013). Their test data comes from a variety of sources, including web pages, bilingual forum posts, and jumbled data from monolingual sources, but none of them are trained on code-switched data, opting instead for a monolingual training set per language. This could prove to be a problem when working on code-switched data, particularly in shorter samples such as social media data, as the code-switching context is not present in training material.

One system tackled both the problems of code-switching and social media in language and code-switched status identification (Lignos and Marcus, 2013). Lignos and Marcus gathered millions of monolingual tweets in both English and Spanish in order to model the two languages, and used crowdsourcing to annotate tens of thousands of Spanish tweets, approximately 11% of which contained code-switched content. This system was able to achieve 96.9% word-level accuracy and a 0.936 F-measure in identifying code-switched tweets.

The issue still stands that relatively little code-switching data, such as that used in Lignos and

Marcus' research, is readily available. Even in their data, the percentage of code-switched tweets was barely over a tenth of the total test data. There have been other corpora built, particularly for other language pairs such as Mandarin-English (Li et al., 2012; Lyu et al., 2010), but the amount of data available and the percentage of code-switching data within that data are not up to the standards of other areas of the natural language processing field. With this in mind, we sought to provide corpora for multiple language pairs, each with a better distribution of code-switching phenomena.

4 Data Sets

Most of the data for the shared task comes form Twitter. However, we also collected and annotated data from other social media sources, including Facebook, web forums, and blogs. These additional sources of data were used as the surprise data. In this section we describe briefly the corpora curated for the shared task.

Language-pair	**Training**	**Test**	Surprise
MAN-EN	1000	313	n/a
MSA-DA	5,838	2332, 1,777	12,017
NEP-EN	9,993	3,018 (2,874)	1,087
SPA-EN	11,400	3,060 (1,626)	1,102

Table 2: Statistics of the shared task data sets per language pairs. The numbers are according to what was actually annotated, numbers in parenthesis show what the participating systems were able to crawl from Twitter. The Surprise genre comes from various sources, other than Twitter.

Table 2 shows some statistics about the different datasets used in this task. We strive to provide dataset sizes that would allow a robust analysis of results. However, an unexpected challenge was the rate at which tweets became unavailable. Different language pairs had different attrition rates with SPA-EN being the most affected language and MSA-DA and NEP-EN the least affected. Note that we provided two test datasets for MSA-DA. Since we separated the data on a per user basis, the first test set had a highly skewed distribution. The second test set was distributed to participants to allow a comparison with a data set having a class distribution more similar to the training set.

4.1 SPA-EN data

Developing the corpus involved two primary steps: locating code-switching tweets and using crowdsourcing to annotate their tokens with language tags. A small portion of the tweets were annotated in-lab and this was used as the gold data for quality control in the crowdsourcing annotation.

To avoid biasing the data used in this task, we used a two step process to select the tweets: first we identified CS tweets by doing a keyword search on Twitter's API. We selected a few frequently used English words and restricted the search to tweets identified by Twitter as Spanish from users in California and Texas. An additional set of tweets was then collected by using frequent Spanish words in an all English tweet, from users in the same locations. We filtered these tweets to remove tweets containing URLs, duplicates, spam tweets and retweets.

In-lab annotators labeled the filtered tweets using the guidelines referenced above. From this set of labeled data we then ranked the users in this set by the percentage of CS tweets. We selected the 12 most prolific CS users and then pulled all of their available tweets. These 12 users contributed the tweets used in the shared task. The tweets were labeled using CrowdFlower[3]. After analyzing the number and content distribution of the tweets, the SPA-EN data was split into a 11,400 tweet training set and a 3,014 tweet test set.

The SPA-EN Surprise Genre (SPA-EN-SG) included Facebook comments from the Veteranas community[4] and the Chicanas community[5] and blog data from the Albino Bean [6]. Data was collected using Python scripts that implemented the Beautiful Soup library and the third-party Python Facebook SDK (for Blogger and Facebook respectively). Post and comment IDs were used to identify Facebook posts, and URLs were used to identify Blogger posts. The collected posts were formatted to match those collected from Twitter. In-lab annotators were used to annotate approximately 1K tokens. All the data we collected in this manner was released as surprise data to all participants.

4.2 NEP-EN data

The collection of NEP-EN data followed a similar approach to that of SPA-EN. We first focused on finding users that switched frequently between

[3] http://www.crowdflower.com/

[4] https://www.facebook.com/VeteranaPinup

[5] https://www.facebook.com/pages/Chicanas/444483772293893

[6] http://thealbinobean.blogspot.com/

Nepali and English. In addition, the users must not be using Devnagari script as done by Nepalese to write Nepali, but must have used its Romanized form. We started by manually reading tweets from some of our Nepali friends. We then crawled their followers who corresponded with them using code-switched tweets or replies. We found that a lot of these users were regular code-switchers themselves. We repeated the same process with the followers and collected nearly 30 such users. We then collected about 2,000 tweets each from these users using the Twitter API. We filtered out all the retweets and the tweets with URLs, following the same process that was used for SPA-EN.

For the surprise test data, we crawled code-switched data from Facebook comments and posts. We found that most Nepalese comments had a rich amount of code-switched data. However, we could not crawl their data because of privacy issues. Nevertheless, we could crawl data from public Facebook pages. We identified some public Nepali Facebook pages where anyone could comment. These pages include FM, news and public figures' public Facebook pages. We crawled the latest 10 feeds from these public pages using the Facebook API and gathered about 12,000 comments and posts for the shared task.

Initially, we sought out help from Nepali graduate students at the University of Alabama at Birmingham to annotate 100 tweets (1739 tokens). We gave the same annotation file to two annotators to do the annotation. We found that they agreed with an accuracy of 95.34%. These tweets were then reviewed and used as initial gold data in Crowdflower to annotate the first 1000 tweets. The annotation job was enabled only in Nepal and Bhutan. We disabled India, even though people living in some regions of India (Darjeeling, Sikkim) also speak and write in Nepali, as most spammers were coming from India. We then ran two batches of 5000 tweets and one batch of 3000 tweets along with the initial 1,000 tweets as the gold data. This NEP-EN data was then split into a 9,993 tweet training set and a 2,874 tweet test set. No Twitter user appeared in both sets.

4.3 MAN-EN data

The MAN-EN tweets were collected from Twitter with the Twitter API. Users were selected from lists of most followed Twitter accounts in Taiwan (where Mandarin Chinese is the official language). These users' tweets were checked for Mandarin English bilingualism and added to our data collection if they contained both languages.

The next round of usernames came from the lists of users that our original top accounts were following. The tweets written by this new set of users were then examined for Mandarin English code switching and stored as data if they matched the criteria.

The jieba tokenizer[7] was used to segment the Mandarin sections of the tweets and compute offsets of each segment. We format the code switching tweets into columns including language type, labels, and offsets. Named entities were labeled manually by a single annotator.

The data was split by user into 1000 tweets for training and 313 for testing. No MAN-EN surprise data for the current shared task.

4.4 MSA-DA data

For the MSA-DA language pair, we selected Egyptian Arabic (EGY) as the Arabic dialect. We harvested data from two social media sources: Twitter [TWT] and Blog commentaries [COM]. The TWT data served as the main gold standard data for the task where we provided fully annotated data for Training/Tuning and Test. We provided two TWT data sets for the test data that exemplified different tag distributions. The COM data set comprised only test data and it served as the Arabic surprise data set.

To reduce the potential of TWT data attrition from users deleting their accounts or tweets, we selected tweets that are less prone to deletion and/or change. Thereby we harvested tweets by a select set of Egyptian Public Figures. The percentage of deleted tweets and deactivated accounts among those users is significantly lower if we compare it to the tweets crawled from random Egyptian users.

We used the "Tweepy" library to crawl the timelines of 12 Public Figures. Similar to other language pairs, we excluded all re-tweets, tweets with URLs, tweets mentioning other users, and tweets containing Latin characters. We accepted 9,947 tweets, for each we extracted the tweet-id and user-id. Using these IDs, we retrieved the tweets text, tokenized it and assigned character offsets. To guarantee consistency and avoid any misalignment issues, we compiled the full pipeline into the "Arabic Tweets Token Assigner" package which is made

[7] https://github.com/fxsjy/jieba

available through the workshop website[8].

For COM, we selected 6723 commentaries (half MSA and half DA) from "youm7"[9] commentaries provided by the Arabic Online Commentary Dataset (Zaidan and Callison-Burch, 2011). The COM data set was processed (12017 total tokens) using the same pipeline created for the task. We also provided the participants with the data formatted with character offsets to maintain consistency across data sets in the Arabic subtask.

The annotation of MSA-DA language pair data is based on two sets of guidelines. The first set is a generic set of guidelines for code switching in general across different language pairs. These guidelines provide the overarching framework for annotating code switched data on the morphological, lexical, syntactic, and pragmatic levels. The second set of guidelines is language pair specific. We created the guidelines for the Arabic language specifically. We enlisted the help of 3 annotators in addition to a super annotator, hence resulting in 4 annotators overall for the whole collection of the data. All the annotators are native speakers of Egyptian Arabic with excellent proficiency in MSA. The super annotator only annotated 10% of the overall data and served as the adjudicator. The annotation process was iterative with several repetitions of the cycle of training, annotation, revision, adjudication until we approached a stable Inter Annotator Agreement (IAA) of over 90% pairwise agreement.

5 Survey of Shared Task Systems

We received submissions from seven different teams. Each participating system had the freedom to submit responses to any of the language pairs covered in the shared task. All seven participants submitted system responses for SPA-EN, making this language pair the most popular in this shared task and MAN-EN the least popular.

All but one participating system used a machine learning algorithm or language models, or even a combination of both, as part of their configuration. A couple of the participating systems used handcrafted rules of some sort, either at the intermediate steps or as the final post-processing step. We also observed a good number of systems using external resources, in the form of labeled monolingual corpora, language specific gazetteers, off the shelf tools (NE recognizers, language id systems, or morphological analyzers) and even unsupervised data crawled from the same users present in the data sets provided. Affixes were also used in some form by different systems.

The architecture of the different systems ranged from a simple approach based on frequencies of character n-grams combined in a rule-based system, to more complex approaches using word embeddings, extended Markov Models, and CRF autoencoders. The majority of the systems that participated in more than one language pair did little to no customization to account for the morphological differences of the specific language pairs beyond language specific parameter-tuning, which probably reflects participants' goal to develop a multilingual id system.

Due to the presence of the NE label, several systems included a component for NE recognition where there was one available for the specific language. In addition, many systems also included case information. One unexpected finding from the shared task was that no participating system tried to embed in their models some form of linguistic theory or framework about CS. Only one system made an explicit reference to CS theories (Chittaranjan et al., 2014) in their motivation to use contextual information, which can be considered as a loose embedding of CS theory. While system performance was competitive (see next section), there is still room for improvement and perhaps some of that improvement can come out of adding this kind of knowledge into the models. Lastly, we were surprised to see that not all systems made use of character encoding information, even though for Mandarin-English that would have been a strong indicator. In Table 3 we present a summary highlighting some of the design choices of participating systems.

6 Results

We used the following evaluation metrics: Accuracy, Precision, Recall, and F-measure. We use F-measure to provide a ranking of systems. In the evaluation at the tweet level we use the standard f-measure. For the evaluation at the token level we use instead the average weighted f-measure to account for the highly imbalanced distribution of classes.

To provide a fair evaluation, we only scored pre-

[8] http://emnlp2014.org/workshops/CodeSwitch/call.html

[9] An Egyptian newspaper, www.youm7.com

System	Machine Learning	Rules	Case	Character Encoding	External Resources	LM	Affixes	Context
(Chittaranjan et al., 2014)	CRF		✔	✔	dbpedia dumps, online sources			± 3
(Shrestha, 2014)		✔		✔	spell checker			
(Jain and Bhat, 2014)	CRF		✔	✔	English dictionary	✔	✔	± 2
(King et al., 2014)	eMM				ANERgazet, TwitterNLP, Stanford NER	✔	✔	✔
(Bar and Dershowitz, 2014)	SVM		✔		Illocution Twitter Lexicon, monolingual corpora (NE lists)	✔	✔	± 2
(Lin et al., 2014)	CRF		✔	✔	Hindi-Nepali Wikipedia, JRC, CoNLL 2003 shared task, lang id predictors: cld2 and ldig		✔	✔
(Barman et al., 2014)	kNN, SVM	✔	✔		BNC, LexNorm		✔	± 1

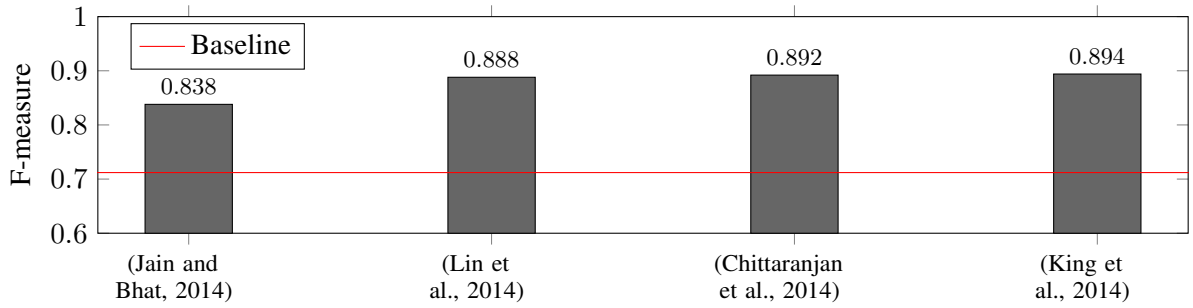
Table 3: Comparison of shared task participating system algorithm choices. CRF stands for Conditional Random Fields, SVM for Support Vector Machines and LM for Language Models.

dictions on tweets submitted by all teams. All systems were compared to a simple lexicon-based baseline. The lexicon was gathered from the training data for classes *lang1* and *lang2* only. Emoticons, punctuation marks, usernames and URLs are by default tagged as *other*. In the case of a tie or a new token, the baseline system assigns the majority class for that language pair.

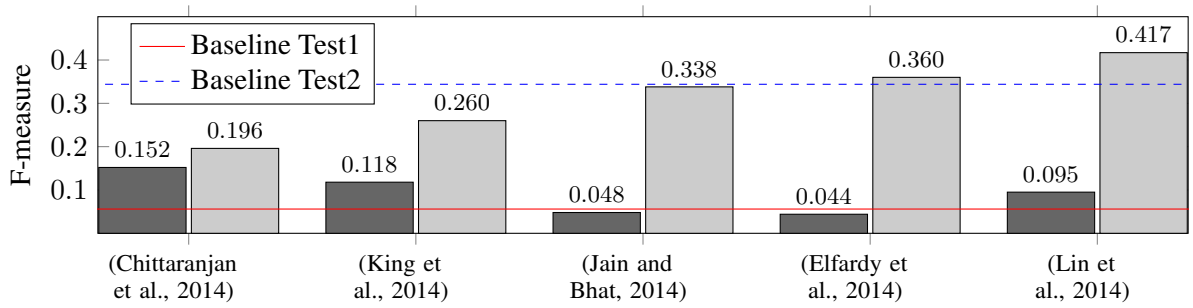
Figure 1 shows prediction performance on the Twitter test data for each language pair at the tweet level. The system predictions for this task are taken directly from the individual token predictions in the following manner: if the system predictions for the same tweet contain at least one tag from each language (lang1 and lang2), the tweet is labeled as code-switched, otherwise it is labeled as monolingual. As illustrated, each language pair shows different patterns. Comparing the systems that participated in all language pairs, there is no clear winner across the board. However, (Chittaranjan et al., 2014) was in the top three places in at least one test file for each language pair. Table 4 shows the results at the token level by label. Here again the figures show F-measure per class label and the last column is the weighted average f-measure (Avg-F). One of the few general trends on these results is that most participating systems were not able to correctly identify the minority classes "ambiguous" and "other". There are only few instances of these labels in the training set and some test sets did not have one of these classes present. The impact on final system performance from these classes is not significant. However, to study CS patterns we will need to have these labels identified properly.

The MAN-EN pair received four system responses and all four of them reached an F-measure $>$80% and outperformed the simple baseline by a considerable margin. We expected this language pair to be the easiest one for the shared task since each language uses a different encoding script. A very rough but accurate distinction between Mandarin and English could be achieved by looking at the character encoding. However, according to the system descriptions provided, not all systems used encoding information. The best performing systems for MAN-EN are (King et al., 2014) and (Chittaranjan et al., 2014). The former slightly outperformed the latter at the Tweet level (see Figure 1a) task while the opposite was true at the token level (see Table 4 rows 4 and 5).

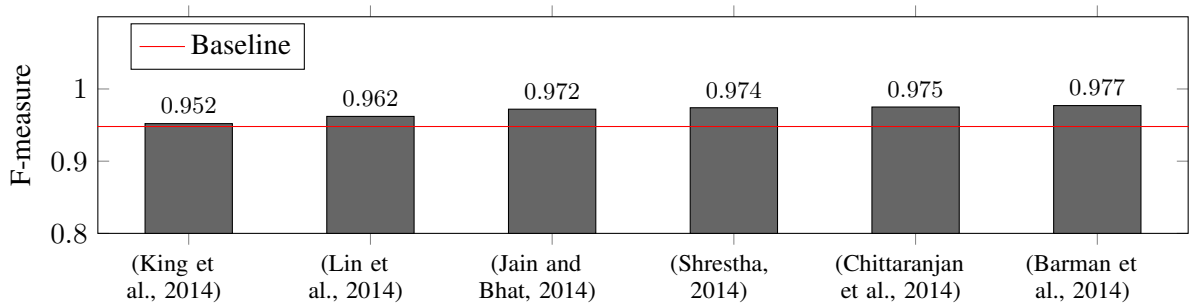
In the case of SPA-EN, all seven systems outperformed the simple baseline. The best performing system in all SPA-EN tasks was (Bar and Dershowitz, 2014). This system achieved an F-measure of 82.2%, 2.9 percentage points above the second best system (Lin et al., 2014) on the tweet level task (see Figure 1(d)). In the token level evaluation, (Bar and Dershowitz, 2014) reached an Avg. F-measure of 94%. This top performing system uses a sequential classification approach where the labels from the preceding words are used as features in the model. Another design choice that might have given the edge to this system is the fact that their model combines character- and word-based language models in what the authors call "intra- and inter-word level" features. Both types of language models are trained on large amounts of monolingual data and NE lists, which again provides additional knowledge that other systems are not exploiting. For instance, the NE lexicons might account for the best results in the NE class in both the Twitter data and the Surprise genre (see Table 4 last row for SPA-EN and second to last for SPA-EN Surprise). Most systems showed considerable



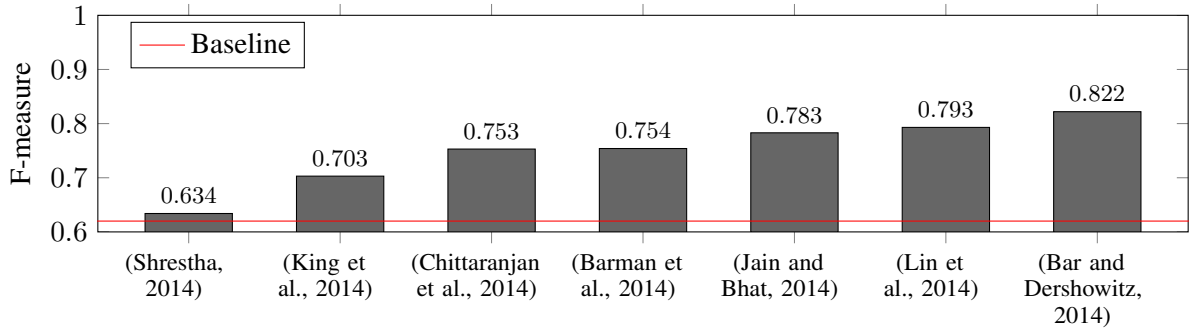
(a) MAN-EN



(b) MSA-DA. Dark gray bars show performance on Test1 and light gray bars show performance for Test2



(c) NEP-EN



(d) SPA-EN

Figure 1: Prediction results on language identification at the tweet level. This is a binary task to distinguish between a monolingual and a CS tweet. We show performance of participating systems using F-measure as the evaluation metric. The solid line shows the lexicon baseline performance.

differences in prediction performance in both genres. In all cases the Avg. F-measure was higher on the Twitter test data than on the surprise genre. Although the surprise genre is too small to draw strong conclusions, all language pairs with surprise genre test data showed a decrease in performance of around 10%.

We analyzed system outputs and found some consistent sources of error. Lexical forms that exist in both languages were frequently mislabeled by

most systems. For example the word for "he" was frequently mislabeled by at least one system. In most of the cases systems were predicting EN as label when the target language was SPA. Cases like this were even more prone to errors when these words fell in the CS point, as in this tweet: *ni el* header ***he*** *hecho* (I haven't even done the header). Tweets like this one, with just one token from the other language, were difficult for most systems. Named entities were also frequent sources of error, especially when they were spelled with lower cases letters.

By far the hardest language pair in this shared task was MSA-DA, as anticipated. Especially when considering the typological similarities between MSA and DA. This is mainly due to the fact that DA and MSA are close variants of one another and hence they share considerable amount of lexical items. The shared lexical items could be simple cognates of one another, or *faux amis* where they are homographs or homophones, but have completely different meaning. Both categories constitute a significant challenge. Accordingly, the baseline system had the lowest performance from all language pairs in both test sets. We note challenges in this language pair on each linguistic level where CS occurs especially for the shared lexical items.

On the phonological level, DA writers tend to mimic the MSA script for DA words even if they are pronounced differently. For example: "heart" is pronounced in DA *Alob* and in MSA as *qalob* but commonly written in MSA as "qalob" in DA data. Also many phonological differences are in short vowels that are underspecified in written Arabic, adding another layer of ambiguity.

On the morphological level, there is no available morphological analyzer able to recognize such shared words and hence they are mostly misclassified. Language identification for MSA-DA CS text highly depends on the context. Typically some Arabic variety word serves as a marker for a context switch such as *mElh$* for DA or *mn∗* for MSA. But if shared lexical items are used, it is challenging to identify the Arabic variant. An example from the training data is *qlb* meaning either *heart* as a noun or *change* as a verb in the phrase *lw qlb mjrm*, corresponding to 'If the heart of a criminal' or 'if he changes into a criminal'. These challenges render language identification for CS MSA-DA data far from solved as evident by the fact that the highest scoring system reached an F-measure of only 41.7% in Test2 for CS identification. Moreover, this is the only language pair where at least one system was not able to outperform the baseline and in the case of Test2 only one system (Lin et al., 2014) outperformed the baseline.

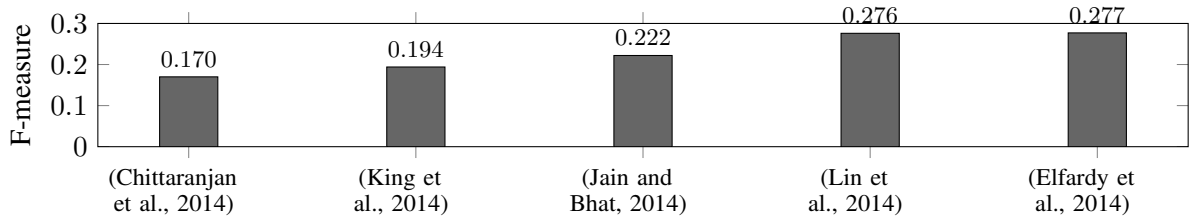
Most teams did well for the NEP-EN shared task, and all teams outperformed the baseline. The reason for the high performance might be the high number of codeswitched tweets in the training and test data for NEP-EN (much higher than other language pairs). This allowed systems to have more samples of CS instances. The other reason for good performance by most participants in both evaluations might be that Nepali and English are two very different languages. The structure of the words and syntax of word formation are very different. We suspect, for instance, that there is a much lower overlap of character n-grams in this language pair than in SPA-EN, which makes for an easier task. At the Tweet level, system performance ranged over a small set of values, the lowest F-measure was 95.2% while the highest was 97.7%. Looking at the numbers in Table 4, we can see that even NE recognition seemed to be a much easier task for this language pair than for SPA-EN (compare results for the NE category in both SPA-EN sets to those of both NEP-EN data sets). The best performing system for the Twitter test data is (Barman et al., 2014) with an F-measure of 97.7%. The results trend in the surprise genre is not consistent with what we observed for the Twitter test data. The top ranked system for Twitter sunk to the 4th place with an F-measure or 59.6%, a considerable drop of almost 40 percentage points. In this case, the overall numbers indicate a much wider difference in the genres than what we observed for other languages, such as SPA-EN, for example. We should note that the class distribution in the surprise data is considerably different from what the models used for training, and from that of the test data as well. In the Twitter data there was a larger number of CS tweets than monolingual ones, while in the surprise genre the majority class was monolingual. This will account for a good portion of the differences in performance. But here as well, the small number of labeled instances makes it hard to draw strong conclusions.

Test Set	System	lang1	lang2	NE	other	ambiguous	mixed	Avg-F
MAN-EN	Baseline	0.9	0.47	0	0.29	-	0	0.761
	(Jain and Bhat, 2014)	0.97	0.66	0.52	0.33	-	0	0.871
	(Lin et al., 2014)	0.98	0.73	0.62	0.34	-	0	0.886
	(King et al., 2014)	0.98	0.74	0.58	0.30	-	0	0.884
	(Chittaranjan et al., 2014)	0.98	0.76	0.66	0.34	-	0	**0.892**
MSA-DA Test 1	(King et al., 2014)	0.88	0.14	0.05	0	0	-	0.720
	Baseline	0.92	0.06	0	0.89	0	-	0.819
	(Chittaranjan et al., 2014)	0.94	0.15	0.57	0.91	0	-	0.898
	(Jain and Bhat, 2014)	0.93	0.05	0.73	0.87	0	-	0.909
	(Lin et al., 2014)	0.94	0.09	0.74	0.98	0	-	0.922
	(Elfardy et al., 2014)*	0.94	0.05	0.85	0.99	0	-	**0.936**
MSA-DA Test 2	Baseline	0.54	0.27	0	0.94	0	0	0.385
	(King et al., 2014)	0.59	0.59	0.13	0.01	0	0	0.477
	(Chittaranjan et al., 2014)	0.58	0.50	0.42	0.43	0.01	0	0.513
	(Jain and Bhat, 2014)	0.62	0.49	0.67	0.75	0	0	0.580
	(Elfardy et al., 2014)*	0.73	0.73	0.91	0.98	0	0.01	0.777
	(Lin et al., 2014)	0.76	0.81	0.73	0.98	0	0	**0.799**
MSA-DA Surprise	(King et al., 2014)	0.48	0.60	0.05	0.02	0	0	0.467
	(Jain and Bhat, 2014)	0.53	0.61	0.62	0.96	0	0	0.626
	(Chittaranjan et al., 2014)	0.56	0.69	0.33	0.96	0	0	0.654
	(Lin et al., 2014)	0.68	0.82	0.61	0.97	0	0	0.778
	(Elfardy et al., 2014)*	0.66	0.81	0.87	0.99	0	0	**0.801**
NEP-EN	Baseline	0.67	0.76	0	0.61	-	0	0.678
	(King et al., 2014)	0.87	0.80	0.51	0.34	-	0.03	0.707
	(Lin et al., 2014)	0.93	0.91	0.49	0.95	-	0.02	0.917
	(Jain and Bhat, 2014)	0.94	0.96	0.52	0.94	-	0	0.942
	(Shrestha, 2014)	0.94	0.96	0.57	0.95	-	0	0.944
	(Chittaranjan et al., 2014)	0.94	0.96	0.45	0.97	-	0	0.948
	(Barman et al., 2014)	0.96	0.97	0.58	0.97	-	0.06	**0.959**
NEP-EN Surprise	(Lin et al., 2014)	0.83	0.73	0.46	0.65	-	-	0.712
	(King et al., 2014)	0.82	0.88	0.43	0.12	-	-	0.761
	(Chittaranjan et al., 2014)	0.78	0.87	0.37	0.80	-	-	0.796
	(Jain and Bhat, 2014)	0.83	0.91	0.50	0.87	-	-	0.850
	(Barman et al., 2014)	0.87	0.90	0.61	0.74	-	-	0.853
	(Shrestha, 2014)	0.85	0.92	0.53	0.78	-	-	**0.855**
SPA-EN	Baseline	0.72	0.56	0	0.75	0	0	0.704
	(Shrestha, 2014)	0.88	0.85	0.35	0.92	0	0	0.873
	(Jain and Bhat, 2014)	0.92	0.92	0.36	0.90	0	0	0.905
	(Lin et al., 2014)	0.93	0.93	0.32	0.91	0.03	0	0.913
	(Barman et al., 2014)	0.93	0.92	0.47	0.93	0.03	0	0.921
	(King et al., 2014)	0.94	0.93	0.54	0.92	0	0	0.923
	(Chittaranjan et al., 2014)	0.94	0.93	0.28	0.95	0	0	0.926
	(Bar and Dershowitz, 2014)	0.95	0.95	0.56	0.94	0.04	0	**0.940**
SPA-EN Surprise	(Shrestha, 2014)	0.80	0.78	0.23	0.81	0	0	0.778
	(Jain and Bhat, 2014)	0.83	0.84	0.22	0.79	0	0	0.811
	(Lin et al., 2014)	0.83	0.86	0.19	0.80	0.03	0	0.816
	(Barman et al., 2014)	0.84	0.85	0.31	0.82	0.03	0	0.823
	(Chittaranjan et al., 2014)	0.94	0.86	0.14	0.83	0	0	0.824
	(King et al., 2014)	0.84	0.85	0.35	0.81	0	0	0.828
	(Bar and Dershowitz, 2014)	0.85	0.87	0.37	0.83	0.03	0	**0.839**

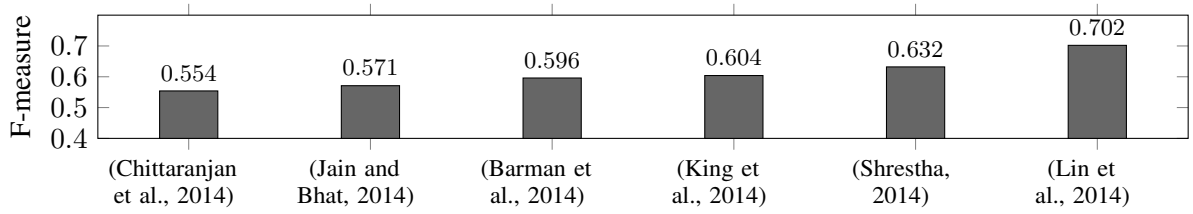
Table 4: Performance results on language identification at the token level. A '-' indicates there were no tokens of this class in the test set. We ranked systems using weighted averaged f-measure (Avg-F). The "*" marks the system by (Elfardy et al., 2014). This system was not considered in the ranking for the shared task as it was developed by co-organizers of the task.

7 Lessons Learned

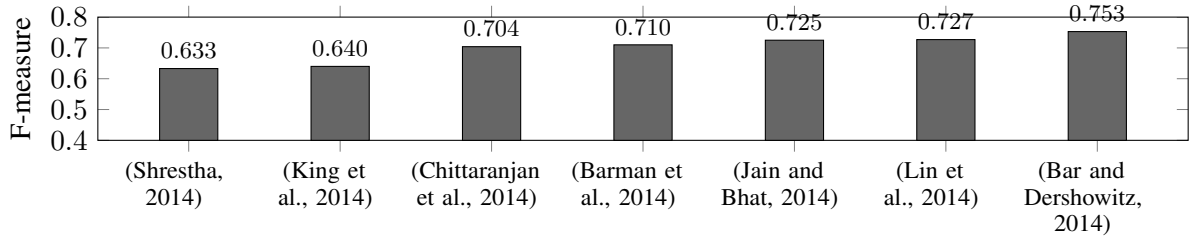
Among the things we want to improve for future shared tasks is the issue of data loss due to removal of tweets or users deleting their accounts. We decided to use Twitter data to have a relevant corpus. However, the trade-off is the lack of rights to distribute the data ourselves. This is not just a burden for the participants. It is an awful waste of resources as the data that was expensive to gather and label is not being used beyond the small group of researchers involved in the creation of the corpus. This will deter us from using Twitter data for future shared tasks, at least until a better solution is identified.



(a) MSA-DA Surprise Genre Results



(b) NEP-EN Surprise Genre Results



(c) SPA-EN Surprise Genre Results

Figure 2: Prediction results on language identification at the document level for the surprise genre. This is a binary task to distinguish between a monolingual and a code-switched text. We show performance of participating systems using F-measure as the evaluation metric.

Using crowdsourcing for annotating the data is a cheap and easy way for generating resources. But we found out that even when following best practices for quality control, there was a substantial amount of noise in the gold data. We plan to continue working on refining the annotation guidelines and quality control processes to reduce the amount of noise in gold annotations.

8 Conclusion

This is the first shared task on language identification in CS data. Yet, the response was quite positive as we received 42 system runs from seven different teams, plus submissions for MSA-AD from a sub group of the task organizers (Elfardy et al., 2014). The systems presented are overall robust and with interesting differences from one another. Although we did not see a single system ranking in the top places across all language pairs and tasks, we did see systems showing robust performance indicating some level of language independence. But the results are not consistent at the tweet/document level. The language pair that proved to be the most difficult for the task was MSA-DA, where the lexicon baseline system was hard to beat even with an F-measure of 47.1%.

This shared task showed that language identification in code-switched data is still an open problem that warrants further investigation. Perhaps in the near future we will see systems that embed some form of linguistic theory about CS and maybe that would result in more accurate predictions.

Our goal is to support new research addressing CS data. Discussions about the challenge for the next shared task are already underway. One possibility might be parsing. We plan to investigate the challenges in parsing CS data and we will start by exploring the hardships in manually annotating CS with syntactic information. We would also like to explore the possibility of classifying CS points according to their socio-pragmatic role.

Acknowledgments

We would like to thank all shared task participants. We also thank Brian Hester and Mohamed Elbadrashiny for their invaluable support in the development of the gold standard data and analysis of results. We also thank the in-lab annotators and the CrowdFlower contributors. This work was partly funded by NSF under awards 1205475 and 1205556.

References

Kfir Bar and Nachum Dershowitz. 2014. Tel Aviv University system description for the code-switching workshop shared task. In *Proceedings of the First Workshop on Computational Approaches to Code-Switching*, Doha, Qatar, October. ACL.

Utsab Barman, Joachim Wagner, Grzegorz Chrupala, and Jennifer Foster. 2014. DCU-UVT: Word-level language classification with code-mixed data. In *Proceedings of the First Workshop on Computational Approaches to Code-Switching*, Doha, Qatar, October. ACL.

Gokul Chittaranjan, Yogarshi Vyas, Kalika Bali, and Monojit Choudhury. 2014. A framework to label code-mixed sentences in social media. In *Proceedings of the First Workshop on Computational Approaches to Code-Switching*, Doha, Qatar, October. ACL.

Heba Elfardy, Mohamed Al-Badrashiny, and Mona Diab. 2014. AIDA: Identifying code switching in informal Arabic text. In *Proceedings of the First Workshop on Computational Approaches to Code-Switching*, Doha, Qatar, October. ACL.

Naman Jain and Riyaz Ahmad Bhat. 2014. Language identification in codeswitching scenario. In *Proceedings of the First Workshop on Computational Approaches to Code-Switching*, Doha, Qatar, October. ACL.

A. Joshi. 1982. Processing of sentences with intrasentential code-switching. In Ján Horecký, editor, *COLING-82*, pages 145–150, Prague, July.

Ben King and Steven Abney. 2013. Labeling the languages of words in mixed-language documents using weakly supervised methods. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1110–1119, Atlanta, Georgia, June. Association for Computational Linguistics.

Levi King, Eric Baucom, Tim Gilmanov, Sandra Kübler, Dan Whyatt, Wolfgang Maier, and Paul Rodrigues. 2014. The IUCL+ system: Word-level language identification via extended Markov models. In *Proceedings of the First Workshop on Computational Approaches to Code-Switching*, Doha, Qatar, October. ACL.

Ying Li, Yue Yu, and Pascale Fung. 2012. A Mandarin-English code-switching corpus. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Mehmet Uğur Doğan, Bente Maegaard, Joseph Mariani, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC-2012)*, pages 2515–2519, Istanbul, Turkey, May. European Language Resources Association (ELRA). ACL Anthology Identifier: L12-1573.

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

Chu-Cheng Lin, Waleed Ammar, Chris Dyer, and Lori Levin. 2014. The CMU submission for the shared task on language identification in code-switched data. In *Proceedings of the First Workshop on Computational Approaches to Code-Switching*, Doha, Qatar, October. ACL.

D.C. Lyu, T.P. Tan, E. Chng, and H. Li. 2010. SEAME: a Mandarin-English code-switching speech corpus in South-East Asia. In *INTERSPEECH*, volume 10, pages 1986–1989.

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, October. Association for Computational Linguistics.

Prajwol Shrestha. 2014. An incremental approach for language identification in codeswitched text. In *Proceedings of the First Workshop on Computational Approaches to Code-Switching*, Doha, Qatar, October. ACL.

Anil Kumar Singh and Jagadeesh Gorla. 2007. Identification of languages and encodings in a multilingual document. In *Proceedings of ACL-SIGWAC's Web As Corpus3*, Belgium.

Omar F. Zaidan and Chris Callison-Burch. 2011. The Arabic online commentary dataset: An annotated dataset of informal Arabic with high dialectal content. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: Short Papers - Volume 2*, HLT '11, pages 37–41, Stroudsburg, PA, USA. Association for Computational Linguistics.