Named Entity Recognition on Code-Switched Data: Overview of the CALCS 2018 Shared Task

Gustavo Aguilar ‡**, Fahad AlGhamdi, Victor Soto** †**,**
Mona Diab, Julia Hirschberg, † and **Thamar Solorio**‡
Department of Computer Science, The George Washington University
{fghamdi, mtdiab}@gwu.edu
‡Department of Computer Science, University of Houston
‡{gaguilaralas, tsolorio}@uh.edu
†Department of Computer Science, Columbia University
†{vs2411, julia}@cs.columbia.edu

Abstract

In the third shared task of the Computational Approaches to Linguistic Code-Switching (CALCS) workshop, we focus on Named Entity Recognition (NER) on code-switched social-media data. We divide the shared task into two competitions based on the English-Spanish (ENG-SPA) and Modern Standard Arabic-Egyptian (MSA-EGY) language pairs. We use Twitter data and 9 entity types to establish a new dataset for code-switched NER benchmarks. In addition to the CS phenomenon, the diversity of the entities and the social media challenges make the task considerably hard to process. As a result, the best scores of the competitions are 63.76% and 71.61% for ENG-SPA and MSA-EGY, respectively. We present the scores of 9 participants and discuss the most common challenges among submissions.

1 Introduction

Code-switching (CS) is a linguistic behavior that occurs on spoken and written language. CS happens when multilingual speakers move back and forth from one language to another in the same discourse. The growing incidence of social media in the way we communicate has also increased the occurrences of code-switching on informal written language. As a result, there is a prevalent demand for more tools and resources that can help to process such phenomenon.

In the previous versions of the Computational Approaches to Linguistic Code-Switching (CALCS) workshop, we focused on providing an annotated corpora for language identification (Solorio et al., 2014; Molina et al., 2016). In this occasion, we extend the annotations to the Named Entity Recognition (NER) level. The goal of this shared task is to provide a code-switched NER dataset that can help to benchmark NER state-of-the-art approaches. This will directly impact the performance of higher-level NLP applications where the code-switching behavior is commonly found.

ENG-SPA Tweet

Original: @_xoxoBecky lmao ni ganas tengo de llorar 😂 , the last movie that made me cry was [***Pineapple Express***]$_{\text{TITLE}}$ 😊 me dejo llorando de risa 😂😂
English: @_xoxoBecky lmao I don't even want to cry 😂 , the last movie that made me cry was [***Pineapple Express***]$_{\text{TITLE}}$ 😊 it left me crying with laughter 😂😂

MSA-EGY Tweet

Buckwalter Encoding: wAy mErkp Dd [***AldAxlyp***]$_{\text{ORG}}$ [***wAmn Aldwlp***]$_{\text{ORG}}$ hbqY sEydp byhA
Arabic: وأي معركة ضد الداخلية وأمن الدولة هبقى سعيدة بيها
English: Any controversy against the Interior Ministry and State Security Service will make me feel happy

Figure 1: Examples of the CALCS 2018 dataset. In the English-Spanish data, the highlighted words represent a movie, tagged as TITLE. While in the MSA-EGY data, the bolded words represent government agencies, tagged as ORGANIZATION

We had a total of 9 participants from which we received 8 submissions on English-Spanish and 5 submissions on Modern Standard Arabic-Egyptian. The best F1-score reported for ENG-SPA[1] was **63.76%** by the **IIT BHU** team (Trivedi et al., 2018) whereas in MSA-EGY[2] was **71.61%**

[1] ENG-SPA competition https://competitions.codalab.org/competitions/18725

[2] MSA-EGY competition https://competitions.codalab.org/competitions/18724

Proceedings of The Third Workshop on Computational Approaches to Code-Switching, pages 138–147
Melbourne, Australia, July 19, 2018. ©2018 Association for Computational Linguistics

by the **FAIR** team (Wang et al., 2018).

2 Task definition

The task consists of recognizing entities in a relatively short code-switched context. The entity types for this task are *person*, *organization*, *location*, *group*, *title*, *product*, *event*, *time*, and *other*. We describe each entity type on Section 3.1. Since NER is a sequential tagging task, we use the IOB scheme to identify multiple words as a single named entity. The addition of this scheme duplicates the number of entities in the task yielding a B(eginning) and I(nside) variations of each of them. This leaves us with 19 possible labels for the classification task.

The evaluation of the task uses two versions of the F1-score. The first is the standard F1, and the second is the Surface Form F1-score introduced by Derczynski et al. (2014). The Surface Form F1-score captures the rare and emerging aspects of the entities. We average both metrics to determine the positions in the leaderboard. Additionally, the shared task was conducted on the CodaLab platform[3], where participants are able to directly evaluate their approaches against the gold data.

3 Datasets

In this section we provide the definition of our labels, describe the annotation process and show the distribution of the ENG-SPA and MSA-EGY datasets.

3.1 Entity instructions

The named entities have been annotated using the instructions below. Note that the definitions of the entity types apply to both language pairs.

- **Person**: This entity type includes proper names and nicknames that can identify a person uniquely. We ignore cases where a person is referred by nouns with adjectives that are not necessarily a nickname. Single artists and famous people are treated as *person*.

- **Organization**: This entity type includes names of companies, institutions and corporations, i.e. every entity that has employees and takes actions as a whole. If the NE can potentially be any other type, the context should be sufficient to support whether it is *organization* or not (e.g., Facebook as organization vs. Facebook as the website application).

- **Location**: This NE refers to physical places that people can visit. It includes cities, countries, addresses, facilities, touristic places, etc. This entity type is not to be confused with *organization*. For instance, when people use organization names to refer to places that can be visited (e.g., restaurants), those entities must be tagged as *location*.

- **Group**: This NE includes sports teams, music bands, duets, etc. *Group* and *organization* are not to be confused. For example, the Houston Astros as a team (i.e., *group*) is different from the Houston Astros institution.

- **Product**: This NE refers to articles that have been manufactured or refined for sale, like devices, medicine, food produced by a company, any well-defined service, website accounts, etc.

- **Title**: This type includes titles of movies, books, TV shows, songs, etc. Very often, titles can be sentences (e.g., the movie *We're the Millers*). *Titles* usually refer to media and must not be confused with the *product* type.

- **Event**: This type refers to situations or scenarios that gather people for a specific purpose such as concerts, competitions, conferences, award events, etc. *Events* do not consider holidays.

- **Time**: This NE includes months, days of the week, seasons, holidays and dates that happen periodically, which are not *events* (e.g., Christmas). It excludes hours, minutes, and seconds. 'Yesterday', 'tomorrow', 'week' and 'year' are not tagged as *time*.

- **Other**: This type includes any other named entity that does not fit in the previous categories. This may include nationalities, languages, music genres, etc.

The motivation behind these entity types partly lies on the contextual difference in which they appear. For instance, when an *organization* can be lexically confused with a *product*, the context should break down the ambiguity. Additionally,

[3] The competitions will be permanently open for future benchmarks

Classes	ENG-SPA			MSA-EGY		
	Train	Dev	Test	Train	Dev	Test
Person	6,226	95	1,888	8,897	1,113	777
Location	4,323	16	803	4,500	474	332
Organization	1,381	10	307	2,596	263	179
Group	1,024	5	153	2,646	303	139
Title	1,980	50	542	2,057	258	18
Product	1,885	21	481	795	81	54
Event	557	6	99	902	121	81
Time	786	9	197	578	79	28
Other	382	7	62	122	19	2
NE Tokens	18,544	219	4,532	23,093	2,711	1,610
O Tokens	614,013	9,364	178,479	181,229	20,031	19,804
Tweets	50,757	832	15,634	10,102	1,122	1,110

Table 1: The named entity distribution of the training, development and testing sets for both language pairs. Note that the *NE tokens* row contains the B(eginning) and I(nside) tokens of the datasets following the IOB scheme. The *O Tokens* row refers to the non-entity tokens.

we tried to include entity types that have an impact on higher-level NLP applications under similar social media scenarios.

3.2 ENG-SPA

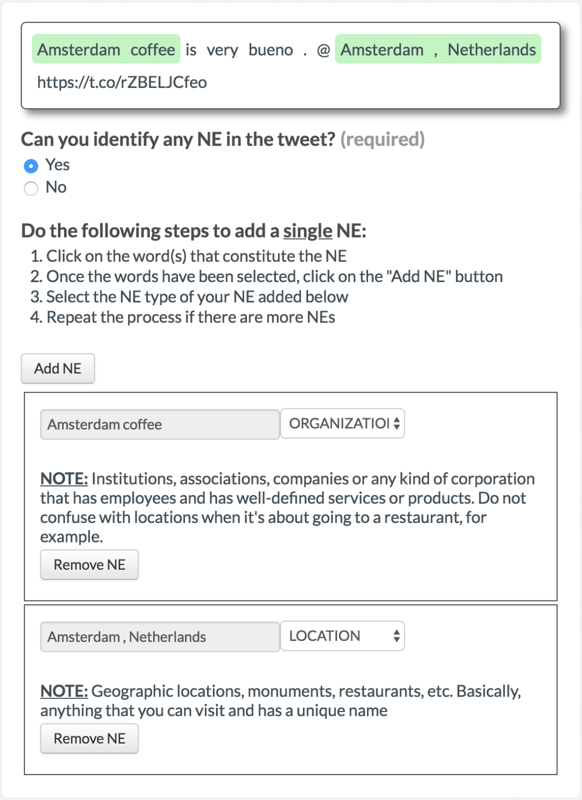
Data annotation: We use the English-Spanish language identification dataset introduced in the first CALCS shared task (Solorio et al., 2014). We build upon this dataset to generate the entity labels. To annotate the data, we designed a CrowdFlower[4] job from scratch[5]. The interface of the job is described in Figure 2. The job allows annotators to select one or many words for a single NE. When the annotators select a word the tool suggests to incorporate words surrounding the current selection. When the selection of a whole entity is done, the annotators can add the entity to the second step where the type is determined. The annotators repeat this process until no more named entities can be identified in the tweet. The output of our customized job contains the entity type of one or multiple words that identify an NE according to the criteria of the annotators. The annotators are required to know both English and Spanish, and the job is constrained to reach an accuracy of at least 80%. We also required 3 annotators per tweet. Additionally, the job was launched in geographic locations were both English and Spanish are reasonably common. Some of these places were USA, Mexico, Central America, Puerto Rico, Colombia, Venezuela, Chile, Uruguay, Paraguay and Spain. After getting the output data from CrowdFlower, we reviewed the results to correct any possible mistakes.

Data distribution: The entity types along with their distribution are listed in Table 1. We provide training, development and testing[6] sets containing 50,757, 832 and 15,634 tweets, respectively. The development and testing splits are inherited from previous CALCS Shared Tasks, whereas training uses the original split with the addition of 40,000 tweets. We added more tweets to the original training set to increase the number of samples per entity type since the NER datasets are naturally skewed. From Table 1, it is worth noting that the total number of NE training tokens is 18,544 whereas the non-entity tokens add up to 614,013. This means that only 3% of the tokens of the training set are NE-related. Likewise, the ratio of tokens for the development and testing sets are 2.3% and 2.5%, respectively. This skewed distribution poses a great challenge considering that the datasets are further separated by 18 fine-grained entity types (i.e., each entity type has a *beginning* and *inside* variations from the IOB scheme). However, we think that the skewness can be reasonably handled with the provided data. Moreover, the training, development and testing sets draw a

[4] `https://crowdflower.com/`

[5] The JavaScript code and HTML/CSS can be found here: `https://github.com/tavo91/ner_annot`

[6] We do not provide the annotations of the test set because we want the CodaLab competition to be used for public benchmark in the future



Amsterdam coffee is very bueno . @ Amsterdam , Netherlands https://t.co/rZBELJCfeo

Can you identify any NE in the tweet? (required)

● Yes
○ No

Do the following steps to add a single NE:

1. Click on the word(s) that constitute the NE
2. Once the words have been selected, click on the "Add NE" button
3. Select the NE type of your NE added below
4. Repeat the process if there are more NEs

Amsterdam coffee — ORGANIZATIOI

NOTE: Institutions, associations, companies or any kind of corporation that has employees and has well-defined services or products. Do not confuse with locations when it's about going to a restaurant, for example.

Amsterdam , Netherlands — LOCATION

NOTE: Geographic locations, monuments, restaurants, etc. Basically, anything that you can visit and has a unique name

Figure 2: The CrowdFlower interface that we developed to annotate the ENG-SPA dataset. The green-highlighted words are the entities selected by the annotator. The words in the same green area describe a single entity. Once the NE selection has been added, the annotators have to select the type of the entities.

very similar data distribution, which can also help to adapt the learning from training to testing.

3.3 MSA-EGY

Validating old tweets: For the Modern Standard Arabic-Egyptian Arabic Dialect (MSA-EGY) language pair, we combined the training, development, and test sets that we used in the EMNLP 2016 CS Shared Task (Molina et al., 2016) to create the new training corpora for the NER Shared Task. The data was harvested from Twitter. We apply a number of quality and validation checks to insure the quality of the old data. Therefore, we retrieved all old tweets using the the new version of the Arabic Tweets Token Assigner which is made available through the Shared Task website [7]. One of the main reasons for the re-crawling step is to eliminate the tweets that have been deleted, or the tweets that belong to the users whose accounts are suspended by Twitter. The other reason is that some tweets may cause encoding issues when they are retrieved using the crawler script. Thus, all these tweets were removed and eliminated. After performing the validation checks, we accepted and published 11,224 tweets (10,102 tweets for the training set, and 1,122 tweets for the development set).

Data creation and annotation: Since we combined the test set used in the EMNLP-2016 CS Shared Task (Molina et al., 2016) with the dataset used in the EMNLP-2014 CS Shared Task (Solorio et al., 2014) to form the new training and development sets, we needed to crawl and annotate a new test set for our new Shared Task. We resorted to using the Tweepy library to harvest the timeline of 12 Egyptian public figures. We applied the same filtration criteria when crawling and building the test set used in the 2016 CS shared task (Molina et al., 2016). We divided the old combined tweets into training and development sets as follows: 80% train set and 10% development set. Thus, we needed ∼ 1,110 tweets, which represents the 10% of the new test set. As we did in the previous Shared Task, we wanted to consider choosing tweets from public figures whose tweets contain more code-switching points. Therefore, we resorted to using the Automatic Identification of Dialectal Arabic (AIDA2) tool (Al-Badrashiny et al., 2015) to perform token-level language identification for the MSA and EGY tokens in context. Public figures with more than 35% of code-switching points in their tweets were considered. The annotation work of the MSA-EGY dataset was done in-lab by two trained Egyptian native speakers. Our annotation team followed the Named Entity Annotation Guidelines for MSA-EGY, which is made available through the Shared Task website [8]. In the two previous editions of the CS Shared Task (Solorio et al., 2014; Molina et al., 2016), we used a Named Entity ("ne") tag. The "ne" tag was defined as a word or multi-word that represents names of a unique entity such as people's names, countries and places, organizations, companies, websites, etc. The AIDA2 tool (Al-Badrashiny et al., 2015) was used to assign initial automatic tags for highly confident data categories

[7] https://code-switching.github.io/2018/

[8] https://codeswitching.github.io/2018/

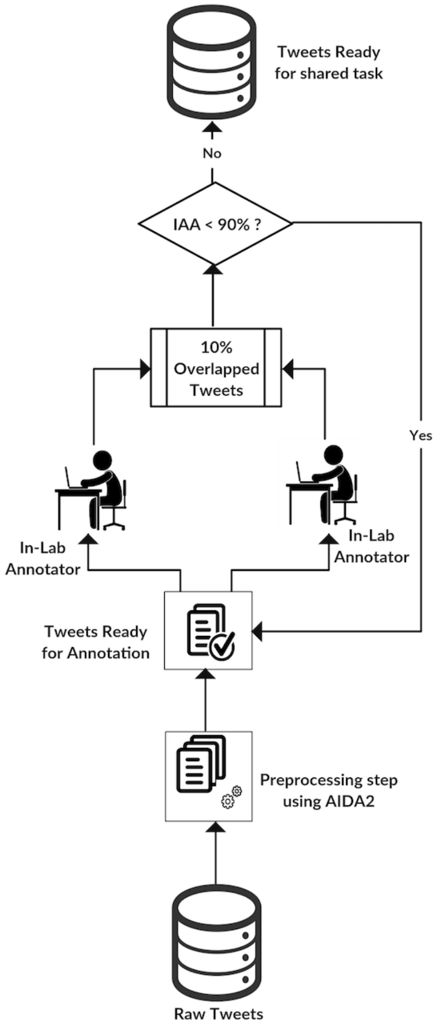


Figure 3: MSA-EGY Data Annotation

(i.e., URL, Punctuation, Number, etc) in addition to named entities. Then, we extracted and prepared all the tweets that contained "ne" for annotation. As we mentioned earlier, the IOB scheme is used as an annotation scheme to identify multiple words as a single named entity. All the URLs, Punctuation and Numbers tags are deterministically converted to "O" tag, while the tweets that include "ne" tags were given to our in-lab annotators for validation and re-annotation if needed.

Quality checks and data distribution: We computed the Inter-Annotator Agreement (IAA) on 10% of the dataset to validate the performance and agreement among annotators. One of our annotators is a specialist linguist who carried out adjudication and revisions of accuracy measurements. We approached a stable Inter Annotator Agreement (IAA) of over 92% pairwise agreement. The workflow of the annotation process for MSA-EGY is shown in Figure-3.

The total number of tweets in MSA-EGY dataset is 12,334 tweets. It is divided into three sets train, development, and test sets (10,102, 1,122, 1,110 tweets, respectively). Table 1 shows that the total number of NE training tokens is 23,093. It means that NE tokens represent 11.3% of the total number of tokens. Similarly, the percentages of NE tokens in the development and test sets are 7.5%, 11.9%, respectively. As we mentioned earlier, the MSA-EGY tweets were harvested from the timeline of 12 Egyptian politicians public figures. Generally, politicians tend to use NEs more often when they write their tweets. This explains why the percentage of the NE tokens in MSA-EGY dataset is higher than the percentage of the NE tokens in ESP-ENG dataset.

4 Approaches

In this section, we briefly describe the systems of the participants and discuss their results as well as the final scores.

- **IIT BHU** (Trivedi et al., 2018). They proposed a "new architecture based on gating of character- and word-based representation of a token". They captured the character and the word representations using a CNN and a bidirectional LSTM, respectively. They also used the Multi-Task Learning on the output layer and transfer the learning to a CRF classifier following Aguilar et al. (2017). Moreover, they fed a gazetteers representation to their model.

- **CAiRE++** (Winata et al., 2018). They used a bidirectional LSTM model for characters and words. They primarily focused on OOV using the FastText library (Bojanowski et al., 2016).

- **FAIR** (Wang et al., 2018). They proposed a joint bidirectional LSTM-CRF network that uses attention at the embedding layer. They also preprocessed the data before feeding the network.

- **Linguists** (Jain et al., 2018). They used a Conditional Random Fields with many handcrafted features. Their focus was primarily on English-Spanish data.

- **Flytxt** (Sikdar et al., 2018). This team also employed a Conditional Random Fields.

Team	Preproc	Ext Res	Hand Feats	CNN	B-LSTM	CRF	Other
IIT BHU		✓		✓	✓	✓	MTL
CAiRE++					✓		FastText
FAIR	✓				✓	✓	Attention
Linguists		✓	✓			✓	
Flytxt		✓				✓	
semantic					✓	✓	
BATs		✓	✓			✓	
Fraunhofer FKIE		✓	✓				SVM
GHHT		✓			✓	✓	

Table 2: The table shows the main component and strategies used by the participants. Ext Res means external resources such as pre-trained word embeddings, gazetteers, etc. Hand Feats means handcrafted features such as capitalization.

They fed the CRF with features from both external and internal resources. Additionally, they incorporated the language identification labels of the datasets from the previous versions of this workshop.

- **semantic** (Geetha et al., 2018). They jointly trained a Bidirectional LSTM with a Conditional Random Fields on the output layer.
- **BATs** (Janke et al., 2018). They used a Conditional Random Fields with multiple features. Some of those features were also used for neural network, but they got better results with the CRF approach.
- **Fraunhofer FKIE** (Claeser et al., 2018). They used a Support Vector Machine (SVM) classifier with a Radial Basis kernel. They handcrafted a lot of features and also included gazetteers.
- **GHHT** (Attia and Samih, 2018). They trained a BLSTM-CRF network using pre-trained word embeddings, brown clusters and gazetteers.
- **Baseline**. We used a simple Bidirectional LSTM network with randomly initialized embedding vectors of 200 dimensions. We also used dropout operations on each direction of the BLSTM component.

5 Evaluation and results

5.1 Evaluation

The evaluation of the shared task was conducted through CodaLab, where the participants were able to obtain immediate feedback of their submissions. The metrics used for the evaluation phase were the standard harmonic mean F1-score and the Surface Form F1 variation proposed by Derczynski et al. (2014). Additionally, to have a single leaderboard per language pair, we unified both metrics by averaging them. The average values are the ones described in Table 3.

As stated by (Derczynski et al., 2014), the idea of the Surface Form F1-score is to capture the *novel* and *emerging* aspects that are usually encountered in social media data. Those aspects describe a fast-moving language that constantly produces new entities challenging more the recall capabilities of state-of-the-art models than the precision side.

5.2 Results and Error analysis

Although all the scores reported by the participants outperformed the baselines in both ENG-SPA and MSA-EGY language pairs, the results are arguably low considering that the current state-of-the-art systems achieve around 91.2% of F1-score on well-formatted text (Lample et al., 2016; Ma and Hovy, 2016; Liu et al., 2017). As mentioned before, the best performing systems reached 63.76% (Trivedi et al., 2018) and 71.61% (Wang et al., 2018) for ENG-SPA and MSA-EGY, respectively. These low outcomes are aligned with the challenges that come along with social media data and the addition of more heterogeneous entity types (Ritter et al., 2011; Augenstein et al., 2017; Derczynski et al., 2014; Aguilar et al., 2018).

Most of the MSA-EGY tweets are related to politics because they were harvested from the

Team	ENG-SPA
IIT BHU	**63.7628**
CAiRE++	62.7608
FAIR	62.6671
Linguists	62.1307
Flytxt	59.2501
semantic	56.7205
BATs	54.1612
Fraunhofer FKIE	53.6514
Baseline	53.2802
	MSA-EGY
FAIR	**71.6154**
GHHT	70.0938
Linguists	67.4419
BATs	65.6207
semantic	65.0276
Baseline	62.7084

Table 3: The results of the participants in both ENG-SPA and MSA-EGY language pairs. The scores are based on the average of the standard and the Surface form F1 metrics. The highlighted teams are the best scores of the shared task.

N	ENG-SPA Samples
1	Retiro totalmente lo dicho sobre **Orange is the New Black**. Temporada terminada y holly sh*t. HOLLY SH*T.
2	**Love Man** by **Otis Redding**, found with @Shazam. Listen now: como me hubiese gustado ver a mis padres bailando esto ...
3	**Michael Jackson** revivió en los **Billboard 2014**
4	@fairy0821 en el **show de shamu** !!!

Table 4: Challenging samples from the test set. The bold words are the ground truth samples and the underscored words are the predictions of the best performing systems.

timeline of number of Egyptian politician public figures. Generally, these kinds of tweets encompass more NEs in comparison with other kinds of tweets. This explains why the percentage of the NE tokens in MSA-EGY dataset is high compared to the NEs' percentage in ESP-ENG data set. This high percentage of NE tokens helps the submitted systems to see and learn more examples and patterns. Thus, systems can generalize more effectively.

According to the results of the participants in the ENG-SPA shared task, the top three most challenging entity types were *event*, *title*, and *time*. It is worth noting that these three classes are more or less the least frequent types in the dataset (see Table 1), which suggests that having more data samples would produce better results. However, in the case of *title*, there are 1,980 samples against 1,381 samples of *organization*, and the performance is significantly better for the latter one (19% vs. 35% of F1-scores). Additionally, looking at Table 4, the entity *Orange is the New Black* was not recognized by participants as a *title*. This is an example of what we refer to heterogeneous entity type, meaning that the entity instances are flexible in format that can even describe independent sentences (i.e., a homogeneous type is *person*). The entities *Love Man* (title), *Billboard 2014* (event), and *show de shamu* (event) also describe the same pattern and they were hardly identified by participants.

Unlike English and Spanish language pair which can be considered as two distinct languages, Modern Standard Arabic and Egyptian are more closely related which makes the task of identifying NE tokens more challenging. This is mainly due to the fact that Modern Standard Arabic and Egyptian are close variants of one another and hence they share considerable amount of lexical items. Some of the challenges faced by the participants include words that still have punctuation attached to them (e.g. مصر) , (mSr, (Egypt) . In order to mitigate these issues, some participants preprocessed these cases by, for example, removing any leading and trailing punctuation from those tokens. Other participants normalized these cases by unifying all the attached punctuations, while the remaining participants decided to keep them and let their model learn them. Table 5 and the following examples show some challenges faced by the submitted systems:

- Clitic attachment can obscure tokens, e.g. والله wAllh "and-God" or "swear".

- Clitic attachment can obscure tokens, e.g. ومنى wmnY "and-Mona" or "swear".

N	MSA-EGY Samples
1	**Buckwalter Encoding:**[*wAllh*]$_{\text{PER}}$ OnA HAss bqhr In [***ElA’ Ebd AlftAH***]$_{\text{PER}}$ [***wmnY***]$_{\text{PER}}$ [***syf***]$_{\text{PER}}$ bytHAkmwA wfy AlqfS **Arabic:** والله أنا حاسس بقهر إن علاء عبد الفتاح ومنى سيف بيتحاكموا وفي القفص **English:** I swear I feel angry knowing that Ala Abdulfatah and-Mona are tried and jailed
2	**Buckwalter Encoding:** kl wAHd ysOl Al—n :[(***mSr***]$_{\text{LOC}}$ rAyHp Ely fyn ?) **Arabic:** كل واحد يسأل الآن : (مصر رايحة علي فين ؟) **English:** Everyone asks himself where is Egypt going to go?

Table 5: Challenging samples from the MSA-EGY test set. The bold words are the ground truth samples.

6 Related work

Before the CALCS workshop series, the code-switching behavior was studied from different perspectives and for many languages (Toribio, 2001; Solorio and Liu, 2008a,b; Piergallini et al., 2016; AlGhamdi et al., 2016). Most of them focused on either exploring this phenomenon or solving core code-switching tasks from the NLP pipeline. More recently, researchers have been considering the sentiment analysis task on code-switching settings (Lee and Wang, 2015; Vilares et al., 2015). However, the lack of resources at the core level of the NLP pipeline greatly reduces the chances of improving higher-level applications. In this line, we aim at providing two datasets for named entity recognition benchmarks on the English-Spanish and Modern Standard Arabic-Egyptian language pairs.

It worth noting that there are some contributions of CS corpora, such as a collection of Turkish-German CS tweets (Calzolari et al., 2016), a large collection of Modern Standrd Arabic and Egyptian Dialectal Arabic CS data (Diab et al., 2016) and a collection of sentiment annotated Spanish-English tweets (Vilares et al., 2016). Named entity recognition has been vastly studied along the years (Sang and Meulder, 2003). More recently, however, the focus has drastically moved to social media data due to the great incidence that social networks have in our daily communication (Ritter et al., 2011; Augenstein et al., 2017). The workshop on Noisy User-generated Text (W-NUT) has been a great effort towards the study of named entity recognition on noisy data. In 2016, the organizers focused on named entities from different topics to evaluate the adaptation of models from one topic to another (Strauss et al., 2016). In 2017, the organizers introduced the Surface Form F1-score metric and collected data from multiple social media platforms (Derczynski et al., 2014). The challenge not only lies on the entity types and the social media noisy but also in the distribution of the datasets and their different data domain patterns.

7 Conclusion

We presented the setup and results of the 3rd shared task of the Computational Approaches to Linguistic Code-Switching workshop. We introduced a named entity recognition dataset focused on code-switched social media text for two language pairs: English-Spanish and Modern Standard Arabic-Egyptian. We received submissions from nine teams, eight of them submitted to ENG-SPA and six to MSA-EGY. Similar to the previous sequence tagging tasks of our workshop, the predominant aspect among the approaches was the Conditional Random Fields. Additionally, the combination of the CRF with a bidirectional LSTM (with some variations) yielded the best results among participants. The best F1-score for ENG-SPA was 63.7628% and for MSA-EGY was 71.6154%. Compared to monolingual formal text (i.e., newswire), the reported scores are significantly lower due to the code-switching phenomenon as well as the noise of SM environment. This serves as strong evidence that we need more robust approaches that can detect and process named entities in such challenging conditions.

Acknowledgments

We would like to thank the National Science Foundation for partially supporting this work under award number 1462142.

References

Gustavo Aguilar, Adrian Pastor Lopez Monroy, Fabio Gonzalez, and Thamar Solorio. 2018. Modeling noisiness to recognize named entities using multitask neural networks on social media. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1401–1412, New Orleans, Louisiana. Association for Computational Linguistics.

Gustavo Aguilar, Suraj Maharjan, Adrian Pastor López Monroy, and Thamar Solorio. 2017. A multitask approach for named entity recognition in social media data. In *Proceedings of the 3rd Workshop on Noisy User-generated Text*, pages 148–153, Copenhagen, Denmark. Association for Computational Linguistics.

Mohamed Al-Badrashiny, Heba Elfardy, and Mona T. Diab. 2015. Aida2: A hybrid approach for token and sentence level dialect identification in arabic. In *Proceedings of the 19th Conference on Computational Natural Language Learning, CoNLL 2015, Beijing, China, July 30-31, 2015*, pages 42–51. ACL.

Fahad AlGhamdi, Giovanni Molina, Mona T. Diab, Thamar Solorio, Abdelati Hawwari, Victor Soto, and Julia Hirschberg. 2016. Part of speech tagging for code switched data. In *Proceedings of the Second Workshop on Computational Approaches to Code Switching@EMNLP 2016, Austin, Texas, USA, November 1, 2016*, pages 98–107.

Mohammed Attia and Younes Samih. 2018. GHHT at CALCS 2018: Named Entity Recognition for Dialectal Arabic Using Neural Networks. In *Proceedings of the Third Workshop on Computational Approaches to Linguistic Code-Switching*, Melbourne, Australia. Association for Computational Linguistics.

Isabelle Augenstein, Leon Derczynski, and Kalina Bontcheva. 2017. Generalisation in named entity recognition: A quantitative analysis. *CoRR*, abs/1701.02877.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2016. Enriching word vectors with subword information. *arXiv preprint arXiv:1607.04606*.

Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Sara Goggi, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asunción Moreno, Jan Odijk, and Stelios Piperidis, editors. 2016. *Proceedings of the Tenth International Conference on Language Resources and Evaluation LREC 2016, Portorož, Slovenia, May 23-28, 2016*. European Language Resources Association (ELRA).

Daniel Claeser, Samantha Kent, and Dennis Felske. 2018. System Description for the Shared Task: Named Entity Recognition on Code-switched Data. In *Proceedings of the Third Workshop on Computational Approaches to Linguistic Code-Switching*, Melbourne, Australia. Association for Computational Linguistics.

Leon Derczynski, Diana Maynard, Giuseppe Rizzo, Marieke van Erp, Genevieve Gorrell, Raphaël Troncy, Johann Petrak, and Kalina Bontcheva. 2014. Analysis of named entity recognition and linking for tweets. *CoRR*, abs/1410.7182.

Mona T. Diab, Mahmoud Ghoneim, Abdelati Hawwari, Fahad AlGhamdi, Nada AlMarwani, and Mohamed Al-Badrashiny. 2016. Creating a large multilayered representational repository of linguistic code switched arabic data. In (Calzolari et al., 2016).

Parvathy Geetha, Khyathi Chandu, and Alan W Black. 2018. Tackling Code-Switched NER: Participation of 'semantic'. In *Proceedings of the Third Workshop on Computational Approaches to Linguistic Code-Switching*, Melbourne, Australia. Association for Computational Linguistics.

Devanshu Jain, Maria Kustikova, Mayank Darbari, Rishabh Gupta, and Stephen Mayhew. 2018. Simple Features for Strong Performance on Named Entity Recognition in Code-Switched Twitter Data. In *Proceedings of the Third Workshop on Computational Approaches to Linguistic Code-Switching*, Melbourne, Australia. Association for Computational Linguistics.

Florian Janke, Tongrui Li, Eric Rincón, Gualberto Guzmán, Barbara Bullock, and Almeida Jacqueline Toribio. 2018. Submission for the Code-Switching Workshop Shared Task 2018 . In *Proceedings of the Third Workshop on Computational Approaches to Linguistic Code-Switching*, Melbourne, Australia. Association for Computational Linguistics.

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural architectures for named entity recognition. *CoRR*, abs/1603.01360.

Sophia Lee and Zhongqing Wang. 2015. Emotion in code-switching texts: Corpus construction and analysis. In *Proceedings of the Eighth SIGHAN Workshop on Chinese Language Processing*, pages 91–99, Beijing, China. Association for Computational Linguistics.

Liyuan Liu, Jingbo Shang, Frank F. Xu, Xiang Ren, Huan Gui, Jian Peng, and Jiawei Han. 2017. Empower sequence labeling with task-aware neural language model. *CoRR*, abs/1709.04109.

Xuezhe Ma and Eduard Hovy. 2016. End-to-end sequence labeling via bi-directional lstm-cnns-crf. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1064–1074, Berlin, Germany. 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.

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*, pages 21–29, Austin, Texas. Association for Computational Linguistics.

Alan Ritter, Sam Clark, Mausam, and Oren Etzioni. 2011. Named entity recognition in tweets: An experimental study. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '11, pages 1524–1534, Stroudsburg, PA, USA. Association for Computational Linguistics.

Erik F. Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the conll-2003 shared task: Language-independent named entity recognition. In *Proceedings of the Seventh Conference on Natural Language Learning, CoNLL 2003, Held in cooperation with HLT-NAACL 2003, Edmonton, Canada, May 31 - June 1, 2003*, pages 142–147.

Utpal Kumar Sikdar, Biswanath Barik, and Björn Gambäck. 2018. Named Entity Recognition on Code-Switched Data using Conditional Random Fields. In *Proceedings of the Third Workshop on Computational Approaches to Linguistic Code-Switching*, 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 and Yang Liu. 2008a. Learning to predict code-switching points. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 973–981, Honolulu, Hawaii. Association for Computational Linguistics.

Thamar Solorio and Yang Liu. 2008b. Part-of-speech tagging for english-spanish code-switched text. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '08, pages 1051–1060, Stroudsburg, PA, USA. Association for Computational Linguistics.

Benjamin Strauss, Bethany Toma, Alan Ritter, Marie-Catherine de Marneffe, and Wei Xu. 2016. Results of the wnut16 named entity recognition shared task. In *Proceedings of the 2nd Workshop on Noisy User-generated Text (WNUT)*, pages 138–144, Osaka, Japan. The COLING 2016 Organizing Committee.

Almeida Jacqueline Toribio. 2001. Accessing bilingual code-switching competence. *International Journal of Bilingualism*, 5(4):403–436.

Shashwat Trivedi, Harsh Rangwani, and Anil Kumar Singh. 2018. IIT (BHU) Submission for the ACL Shared Task on Named Entity Recognition on Code-switched Data. In *Proceedings of the Third Workshop on Computational Approaches to Linguistic Code-Switching*, Melbourne, Australia. Association for Computational Linguistics.

David Vilares, Miguel A. Alonso, and Carlos Gómez-Rodríguez. 2015. Sentiment analysis on monolingual, multilingual and code-switching twitter corpora. In *Proceedings of the 6th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 2–8. Association for Computational Linguistics.

David Vilares, Miguel A. Alonso, and Carlos Gómez-Rodríguez. 2016. En-es-cs: An english-spanish code-switching twitter corpus for multilingual sentiment analysis. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France. European Language Resources Association (ELRA).

Changhan Wang, Kyunghyun Cho, and Douwe Kiela. 2018. Code-Switched Named Entity Recognition with Embedding Attention. In *Proceedings of the Third Workshop on Computational Approaches to Linguistic Code-Switching*, Melbourne, Australia. Association for Computational Linguistics.

Genta Indra Winata, Chien-Sheng Wu, Andrea Madotto, and Pascale Fung. 2018. Bilingual Character Representation for Efficiently Addressing Out-of-Vocabulary in Code-Switching Named Entity Recognition. In *Proceedings of the Third Workshop on Computational Approaches to Linguistic Code-Switching*, Melbourne, Australia. Association for Computational Linguistics.