Token Level Identification of Linguistic Code Switching

Heba Elfardy[1] *Mona Diab*[2]
(1) Department of Computer Science, Columbia University, New York, NY
(2) Center for Computational Learning Systems, Columbia University, New York, NY
heba@cs.columbia.edu, mdiab@ccls.columbia.edu

Abstract

Typically native speakers of Arabic mix dialectal Arabic and Modern Standard Arabic in the same utterance. This phenomenon is known as linguistic code switching (LCS). It is a very challenging task to identify these LCS points in written text where we don't have an accompanying speech signal. In this paper, we address automatic identification of LCS points in Arabic social media text by identifying token level dialectal words. We present an unsupervised approach that employs a set of dictionaries, sound-change rules, and language models to tackle this problem. We tune and test the performance of our approach against human-annotated Egyptian and Levantine discussion fora datasets. Two types of annotations on the token level are obtained for each dataset: context sensitive and context insensitive annotation. We achieve a token level $F_{\beta=1}$ score of 74% and 72.4% on the context-sensitive development and test datasets, respectively. On the context insensitive annotated data, we achieve a token level $F_{\beta=1}$ score of 84.4% and 84.9% on the development and test datasets, respectively.

Keywords: Linguistic Code Switching, Dialect Identification, Modern Standard Arabic, Dialectal Arabic, Dictionaries, Language Models, Sound Change Rules.

Title and Abstract in Arabic:

تحديد نقاط التحول اللغوي على مستوي مفردات الجملة

عادة ما يمزج ناطقوا اللغة العربية بين الفصحى والعامية او اللهجات المختلفة في نفس الجملة وتعرف هذه الظاهرة بالتحول اللغوي. ويمثل تحديد نقاط التحول اللغوي تحدي بسبب عدم وجود الاشارات الصوتية الدالة علي اللهجة. في هذا البحث نهدف الى التعرف الياً على نقاط التحول اللغوي في النصوص من خلال تحديد لهجات الكلمات او مفردات الجملة. ولمعالجة هذه المشكلة نستخدم مجموعة من القواميس بالاضافة الى نماذج اللغة وقواعد لحصر تشابه الاصوات. وقد اختبرنا عدة اعدادات للنظام المقترح على مجموعتين من البيانات المرقمة لغوياً. الاولي تحدّد الفئة الخاصة بالكلمة بغض النظر عن السياق في حين ان الثانية تاخذ السياق في الاعتبار. وحقق النظام المقدم في هذا البحث معدلات وصلت الى ٧٤ % على مجموعة البيانات الخاصة بالظبط و ٧٢،٤ % على البيانات الخاصة بالاختبار في البيانات المعتمدة على السياق. و ٨٤،٤ % و ٨٤،٩ % على مجموعتي البيانات الغير معتمدة على السياق، تباعاً .

التحول اللغوي، تحديد اللهجة، اللغة العربية الفصحى، اللهجات العربية، نماذج اللغة، قواعد تغيّر الصوت

Proceedings of COLING 2012: Posters, pages 287–296,
COLING 2012, Mumbai, December 2012.

1 Introduction

Linguistic Code Switching (LCS) refers to the phenomenon where speakers switch between multiple languages within the same utterance (intra-utterance) or across utterances within the same conversation (inter-utterance). For an example of intra-utterance LCS, consider "Starting a sentence in English, *mais je finis* the same sentence *en Français*", where the italicized words are in French meaning 'but I finish... in French'. Intra-utterance LCS poses a significant challenge for language technologies since ideally one would need to use language processing for both languages simultaneously. In this paper we are mostly interested in intra-utterance LCS. Techniques trained for one language quickly break down when there is input from another. Intra-utterance LCS is quite pervasive in bilingual communities but it is quite pronounced in diglossic languages (Ferguson, 1959) where two forms of the language live side by side and are closely related. This is the case for Arabic where the official form of the language Modern Standard Arabic (MSA) and the dialects (DA), corresponding to the native tongue of the speakers of Arabic, are frequently used together within the same utterances/sentences. There are significant linguistic differences between MSA and DA phonologically, morphologically, lexically and syntactically; MSA is the only standardized written form of the language hence people have no standards for writing DA; and there is a pervasive presence of faux amis between MSA and DA, where words look the same (homographs or homophones) but have different semantic and pragmatic connotations. These differences lead to an exacerbation of the challenges posed by LCS – due to its pervasiveness – on processing informal textual Arabic sources such as news groups, tweets, blogs, and other social media, which are increasingly being studied as rich sources of social, commercial and political information. In this paper, we tackle the problem of identifying LCS points on the token level in a given Arabic text. We cast the problem as a token level dialect identification problem. We incorporate a variety of resources including dictionaries and language models to automatically identify the dialect id of a word in context. We adopt a classification perspective on the problem, hence each token is labeled with a class id (MSA/DA/UNKNOWN). We tune and test different settings of the system. Our approach also allows for producing MSA equivalents and English glosses for the identified DA words. Identifying the classes and sequences of MSA vs. DA words in an utterance can allow for better modeling of Arabic language usage and processing. Moreover the dialect id component can be used for smart filtering for various levels of domain adaptation and targeted document search in an Information Retrieval framework in a rapid process of identifying whether a document is predominantly MSA or DA.

2 Related Work

While there has been considerable interest in LCS from the theoretical and socio-linguistic communities, there has, with few exceptions (Joshi, 1985) (Chan et al., 2004), (Solorio and Liu, 2008a), (Solorio and Liu, 2008b) and (Manandise and Gdaniec, 2011), been little research in computational approaches to the problem. Predictive models of how and when LCS typically occurs, as well as how to interpret LCS items in the context of the matrix language, have yet to be developed. A major barrier to research on LCS has been the lack of large, consistently and accurately annotated corpora of LCS data. In fact, there has been very little discussion even of how such data should be collected and annotated to best support the interests of both the theoretical and the computational communities. (Diab and Kamboj, 2011), and (Elfardy and Diab, 2012) attempted to tackle this problem by annotating corpora of Hindi-English and MSA-DA code switched social media text.
For Chinese English LCS, (Lyu et al., 2006) found that building a unified acoustic model of the regional dialects to be detected, a bilingual pronunciation model, and a Chinese character-based tree-structured search strategy improved ASR performance significantly. For Spanish-English LCS,

Input	dh [1]	Al~y	byHSl	fy	Alwqt	AlrAhn
Eng-GL	that	what	-	in	time	current
MSA-GL	*lk	Al*y	-	fy	Alwqt	AlrAhn
No-Context	DA	DA	DA	MSA-DA	MSA-DA	MSA
Contextual	DA	DA	DA	MSA	MSA	MSA

Table 1: An example of the output of AIDA.

(Solorio and Liu, 2008b) found that LCS poses a serious challenge to part-of-speech tagging: while monolingual taggers reach >96% accuracy, English taggers tested on Spanish-English LCS data obtain only 65% accuracy. Moreover, (Manandise and Gdaniec, 2011) analyzed the effect on Machine Translation quality of borrowing and LCS of Spanish-English within the context of IBM's *"TranslateNow!"* email system. Their study showed that borrowing and LCS degrade the performance of the syntactic parser because these switched tokens are mostly treated as nouns, which results in erroneous analysis and in some cases incomplete parses. As mentioned earlier LCS is even more prominent in Arabic due to the *diglossic* nature of the language yet most of the research effort carried out to tackle Arabic NLP focuses on MSA. A significant exception in Arabic speech processing is work by (Biadsy et al., 2009). In this work, (Biadsy et al., 2009) present a system that identifies dialectal words in speech and their dialect of origin.

3 Approach

In this paper, we present a system, AIDA (Automatic Identification of Dialectal Arabic), that incorporates a set of Language Models, Dictionaries, MSA Morphological Analyzer and Sound-Change-Rules in order to perform Token-Level Dialect Identification. Table 1 shows a sample of the output produced when applying AIDA on a sample Arabic sentence that exhibits LCS. Two outputs are produced for each word in the given sentence. The first of which is a *context-insensitive* output while the second is a *context-sensitive* one. Moreover, AIDA also yields word level glossing in both English and MSA for the DA words.

3.1 Pre-processing

Both corpora used for language models and input text undergo a simple cleaning step. The cleaning process prunes out noisy data yet maintaining all the signals that can help in identifying DA content. This cleaning step: separates punctuation and numbers from words; handles speech effects such as 'goaaaaaaaal', it reduces it to 'goaaal', hence reducing the repeated characters to a maximum of three consecutive repeated characters thereby normalizing all the occurrences of these words to the same form but also maintaining the information that there is a speech effect – a potential clue to dialectalness. This module assigns tokens that have a speech effect a *speech-effect-score*; We also map Latin words, URLs, digits, and punctuation to LAT, URL, NUM, and PUNC class labels, respectively.

For the current implementation of AIDA, we only focus on Arabic written in Arabic script, hence we do not address the problem of romanized Arabic writing. Therefore, any text written using Latin script is replaced by the token LAT which could in principle include romanized Arabic.

In general written Arabic is underspecified for short vowels and consonantal gemination markers which are expressed via diacritics. We find more diacritized words in MSA text than in DA text. In social-media text, we rarely observe the use of diacritics except occasionally for MSA. Therefore, we remove diacritics from all the tokens (LM corpora, input text, and dictionaries) so as to reduce

[1] We use Buckwalter transliterated Arabic. www.qamus.org/transliteration.htm

the variation in forms of the tokens, thereby reducing sparseness. However we assign each token a *diacritization-score* based on the percentage of diacritics it had in the raw text.

3.2 Dialectal Dictionaries

For the DA data we use machine readable dictionaries (MRDs) that are developed for the system Tharwa (Diab et al., 2013). The dictionary, Tharwa, is a three way DA-MSA-English MRD. Tharwa is based on paper dictionaries combined with other resources obtained from the Linguistic Data Consortium (LDC). Tharwa comprises DA lemmas, some surface forms and their corresponding MSA and English equivalents. We have two DA dictionaries: Egyptian and Other-Dialects (mostly Lebanese and Iraqi Arabic). The *Egyptian* Dictionary comprises 33,955 unique DA entries; and the *Other-Dialects* Dictionary comprises 6,926 unique DA entries.[2] At this point we are not addressing Word Sense Disambiguation, hence we merged all the different senses of each word in one entry. However to improve the output of the MSA and English Equivalents for given tokens, the MSA equivalents (Lemmas) are sorted by their frequency of occurrence in the Arabic Gigaword (AGW4).[3]

3.3 ALMOR

We are interested in knowing if a token is MSA or not. We employ a system of MSA morphological analysis, ALMORGEANA (ALMOR) (Habash, 2007). ALMOR relies on the LDC SAMA (Maamouri et al., 2010) database to generate the list of all possible morphological analyses for a word out of context. Moreover, ALMOR provides the English glosses for the analyzed words. If a word has an analysis according to ALMOR, we assume it is MSA.[4] If an analysis is found and it doesn't belong to a predefined DA list then the word is assumed MSA and assigned a score of 1. If the word is analyzed by ALMOR and it belongs to the dialectal-entries' list we assume it is DA and it is assigned a score of 0.5. We limit the number of produced English glosses by having the internal MSA SAMA database entries ranked by their frequency of occurrence in the AGW4.

3.3.1 Using Sound Change Rules for OOVs

If the word isn't successfully analyzed by ALMOR and is not in our DA dictionaries, we attempt a relaxed match on the token using sound change rules (SCR) that model the possible phonological variants of the token. We use a subset of the SCR proposed by (Dasigi and Diab, 2011). Table 2 shows the SCR used. In this case, if the relaxed approximated phonological variant of the word is found by ALMOR, we tag the input word as DA not MSA, and assign it a DA score of 0.5; *and not 1 since the word might be a misspelled MSA word and not a DA variant*; but return the MSA relaxed variant as the MSA equivalent and the corresponding English gloss.

3.4 Language Models (LM)

3.4.1 Data Collection

Our data collection comprises various genres. For the MSA-LM we used a subset of the Arabic Gigaword (AGW4) (Parker et al., 2009), Broadcast News, Broadcast Conversations, and Web-Logs obtained from LDC as well a subset of a more formal MSA-corpus produced by (Rashwan et al.,

[2]The number of entries in these dictionaries reflects the number of undiacritized types (and not tokens) in the original sources.

[3]Detailed information about the dictionaries and their content can be found in (Diab et al., 2013)

[4]Out of the 42,334 lemma entries in the SAMA database, we manually identified 1,725 DA entries. Some of these DA entries could be found in MSA but with extremely low probability.

Letter(s)	Variants	Letter(s)	Variants	Letter(s)	Variants	Letter(s)	Variants
{ < > ' } |	A	t	T v	E	H	g	E x
v	s t S	j	q y $	H	h E	d	∗ D
∗	d z Z	z	∗ Z d	s	$ S v	$	s v
S	s	D	Z d z ∗	T	S Z t	Z	T D z d ∗

Table 2: SCR rules used to expand the coverage of the MSA morphological Analyzer.

2011). We create a small highly dialectal lexicon of words that can rarely or never be used in MSA, we use it to filter out sentences from the MSA corpora thereby attempting to have a more homogeneously MSA collection.
For the DA-LM we use DA news-articles, users-commentaries, DA speech-transcriptions, DA wikipedia, DA poems as well as DA web-logs.
All the corpora undergo the same cleaning preprocessing as described in Subsection 3.1. The corpora DA and MSA comprise 13M tokens each.

3.4.2 Building the Language Models

We use the SRILM toolkit (Stolcke, 2002). We build two 3-gram LMs; (1) *MSA-LM* and (2) *DA-LM* using Kneser-Ney discounting. Using both LMs with the Mix-LM capability in SRILM we create a mixture LM, we allow equal weights for both LMs, thereby creating a third LM, *MSA-DA LM*, that incorporates the entries in both LMs.

3.4.3 Dialect Identification Using LM

From the DA-LM and MSA-LM we build three lists of n-grams, (1) Shared-MSA-DA, (2) MSA-Unique, and (3) DA-Unique Shared-MSA-DA contains the n-grams that are shared between the MSA and DA LMs, while the MSA-Unique and DA-Unique contain entries that exist only in either the MSA-LM or the DA-LM, respectively.
For the shared n-gram list each entry lists: (1) the n-gram, (2) its probability in the MSA-LM, and (3) its probability in the DA-LM. Using these probabilities, we rank the n-gram in each list, the higher the probability, the lower the rank. We then calculate the DA and MSA scores of each n-gram as follows:

$$MSA\_Score_1 = 1 - (MSA\_Rank/Size(Shared\_n-grams\_List))$$
$$DA\_Score_1 = 1 - (DA\_Rank/Size(Shared\_n-grams\_List))$$

We run each input sentence through the mixed-language model in order to divide the sentence into a set of n-grams. For each of the resulting n-grams we check whether it belongs to the Shared-MSA-DA, MSA-Unique or DA-Unique n-gram list.
If the n-gram belongs to the MSA-Unique list, each token in the given n-gram is assigned an MSA score of 1 and a DA score of 0. Conversely if it belongs to the DA-Unique list, then the n-gram tokens are assigned a DA score of 1 and and MSA score of 0.
When the n-gram belongs to the Shared-MSA-DA list, we calculate the difference between the MSA_Score_1 and DA_Score_1 of the n-gram. If the difference is above a certain threshold, we maintain the previously calculated scores, otherwise we update the MSA and DA scores as follows:

$$MSA\_Score_2 = DA\_Score_2 = Maximum(MSA\_Score_1, DA\_Score_1)$$

We experimented with different thresholds (0, 0.1, 0.2, ..., 0.9) on the development (tuning) dataset and got the best results with *0.4* and *0* for the context-insensitive and context-sensitive datasets, respectively.

4 Experiments and Results

We carried out five experimental conditions using the different resources for Dialect Identification.

4.1 Evaluation Dataset

Our approach is unsupervised hence we only annotated data for development and evaluation. We harvested the data from Egyptian and Levantine fora yet there was a significant number of Gulf posts. We annotated 1,170 forum posts corresponding to a total of 27,173 tokens; excluding punctuation, numbers and tokens written in romanized script, yielding 11,767 types. Half of the data comes from Egyptian fora while the other half comes from Levantine ones. Moreover, the data is chosen in a way so as to balance the DA and MSA content. We annotated the data in two different ways: on the word level without much attention to the context (context-insensitive), and contextually where the class of the word highly depends on the context of the text it occurred in (context-sensitive).
Context-Sensitive/Contextual Annotation The annotators are asked to consider the word in context and read it out aloud to themselves to make a decision on whether the word is deemed MSA or DA. For example If a word is used in both MSA and DA with the same sense but occurs in a DA context then it is deemed DA.
Context-Insensitive/No-Context Annotation The annotators perform a per-word annotation meaning that if a word is used in MSA and DA with the same sense then it is assigned a class-label of "MSA-DA".
The Contextual Annotation is more useful in evaluating how well our system is doing on detecting code-switch points while the No-Context helps in assessing the coverage of our MSA and DA resources. For our experiments we split both the Egyptian and Levantine datasets into development and test sets independently. We then merge the development sets from both dialects together and do the same for the test sets, resulting in an Egyptian-Levantine development set and an Egyptian-Levantine test set.

4.2 Dialect Identification Results

We have five Dialect Identification (DID) experimental conditions. Below is a detailed description of how we calculated the score of each token in each of the five experimental set ups. Table 3 shows the results obtained using each of these conditions on the no-context and contextual datasets. We exclude all tokens that are labeled *Named-Entities* or *Foreign* from the evaluation process and consider all tokens labeled as Typos to be *Unknown* words. For all experiments we initialize the DA-score to 1 if the word has consecutive repeated characters (speech-effects) and 0 otherwise, and for the MSA-Score initialization we are guided by the diacritics-scores as described earlier.

DID-1: (Using DICTs and ALMOR) We calculate the MSA score based on analysis retrieved by ALMOR and the DA score from both the dialectal dictionaries and ALMOR (as described earlier, recall that we identified the dialectal entries in the underlying dictionary used by ALMOR and assigned these entries a DA score as opposed to an MSA score). The two scores for DA are then summed and the class of the given token is chosen based on comparing the scores of the two class labels: MSA vs. DA.
DID-2: (Using DICTs, ALMOR and SCR) We use the DA Dictionaries, and attempt to increase the coverage of ALMOR based on Sound Change Rules (SCR). Scores for words that are identified using SCR relaxation are calculated using the approach described earlier (See subsection 3.3.1) and again the scores for the different components are summed prior to identifying the class of the token of interest.

	Dev No-Context						Dev Contextual					
	BL	1	2	3	4	5	BL	1	2	3	4	5
MSA	72.1	92.3	92.3	84.2	88.0	80.8	62.5	81.7	81.7	77.8	82.0	82.0
DA	72.1	62.3	64.6	87.7	73.6	73.9	48.5	45.6	49.5	62.6	64.1	64.6
UNK	2.1	21.2	22.0	18.1	26.2	23.3	2.1	21.2	22.0	18.1	26.2	23.3
All	71.4	77.2	78.6	**84.4**	80.4	80.8	55.6	66.3	68.0	68.5	73.8	**74.0**
	Test No-Context						Test Contextual					
	BL	1	2	3	4	5	BL	1	2	3	4	5
MSA	73.4	92.8	92.8	83.5	87.9	88.1	60.0	75.3	75.3	74.3	77.9	77.8
DA	72.6	62.7	64.3	88.8	74.0	74.3	52.1	50.3	52.7	67.0	66.4	66.7
UNK	0.0	16.7	17.7	14.2	24.6	22.5	0.0	16.7	17.7	14.2	24.6	22.5
All	72.6	78.3	79.3	**84.9**	80.8	81.1	55.8	64.1	65.3	68.8	72.2	**72.4**

Table 3: Token based $F_{\beta=1}$ scores of a random-baseline and the different experimental-conditions on both the context-insensitive and context-sensitive development and test datasets.

DID-3: (Using LMs only) In this condition we assign the score to each token based on the approach described in subsection 3.4.
DID-4: (Using DICTs, ALMOR, and LMs) In this condition, we combine the LMs, DA dictionaries and ALMOR scores.
DID-5: (Using DICTs, ALMOR, LMs, and SCR) In this condition we combine all the scores from all resources and the class for the word is based on the highest aggregate score per class We also calculate a random baseline (BL). We report all our results $F_{\beta=1}$ score metric.

5 Discussion

All the experimental conditions significantly beat the baseline BL. The language-model based approach (DID-3) yields better results than the Dictionary-based and hybrid conditions (DID-1, DID-2, DID-4, and DID-5) on the no-context dataset. Because currently we only use an MSA morphological analyzer (and not a DA one), the dictionary-based and hybrid approaches will bias the predicted class of "MSA-DA" surface tokens towards MSA. ALMOR will produce correct analyses for these tokens while the DA dictionaries won't be able to identify them due to lack of coverage of the different morphological forms – most of our DA entries in the Tharwa dictionary are lemmas – and moreover the problem is exacerbated by the inherent orthographic variance in the DA data yielding potential differences between the data used in the LM and the input data. An example of this is the word *"mdrsthm"* which means *"their school"*, that won't be identified by our DA dictionaries because of the inflection but will be identified by ALMOR.
On the other hand, the hybrid approach performs better on the contextual annotation since we have very few "MSA-DA" tokens in this case hence biasing the system towards choosing only one label is desirable.
While adding the SCR component always yields better results, the absolute magnitude of improvement is diminished when using SCR with LM since LMs increases the coverage of DA words. However SCR are still very useful in getting the MSA-Equivalent of a DA word without having to add more entries to the DA dictionary.
The percentage of OOVs (words that were unrecognized by our system) are much less on MSA tokens compared to DA tokens in the contextual case. The better performance on the MSA data is again attributed to the use of the MSA morphological analyzer that gives better coverage on surface form MSA words; a capability that we currently don't have for DA.
Table 4 shows the details of the confusability between different classes for the best experimental

conditions on the no-context and contextual Test-datasets respectively.

No-Context					Contextual				
	P-MSA	P-DA	P-UNK	A-Tot.[5]		P-MSA	P-DA	P-UNK	G-Tot.
G-MSA	7818	1878	433	10190	G-MSA	5907	2176	14	6833
G-DA	634	9560	389	10036	G-DA	2385	3839	148	5413
G-UNK	52	40	61	153	G-UNK	45	68	40	153

Table 4: Confusion matrix for MSA, DA and UNK classes of Test for conditions that yielded best results. (DID-3 for the context-insensitive dataset and DID-5 for the context-sensitive dataset). The G-MSA/G-DA/G-UNK correspond to gold manual labels while P-MSA/P-DA/P-UNK correspond to the predicted labels (AIDA output)

Context-Insensitive [DID-3] For MSA words, we note that 18.4% of the words are confused for being DA while only 4.2% of the MSA words are classified as UNK reflecting the high-coverage level of our LMs. For DA words, we note that 6.3% of the words are misclassified as MSA and 3.9% of the DA words are classified as UNK. In general, this indicates that we have good coverage DA corpora for LM but more importantly it suggests that our MSA LMs include a residual significant amount of DA data.

Context-Sensitive [DID-5] For MSA words, we note that a significant percentage (31.8%) of the words are confused for being DA. A tiny percentage is confused for being UNK (0.2%). For DA words, we note that a similarly significant percentage, 44.1% of the words, are misclassified as MSA and 2.7% of the DA words are classified as UNK. It does make sense due to the nature of the data since the conditions of both MSA and DA are hard to tell apart. In the contextual annotation guidelines we almost force the annotator to choose between DA or MSA allowing for a "MSA-DA" interpretation only when there isn't enough context (mostly in extremely short phrases). The overall numbers indicate that DA was much harder to classify than MSA words.

Similar to the context-insensitive annotation condition, the majority of the UNK are classified as MSA. In general compared to the results of the context-insensitive condition confusion matrix, we note that there seems to be significantly more confusion among the classes for the contextual conditions.

6 Conclusion[6]

In this paper, we presented several combinations of resources to address the problem of automatic identification of token level dialectalness. The resources include Dictionaries, Morphological Analyzer, Sound Change Rules and Language Models . We evaluate the system performance against forum data pertaining to Egyptian and Levantine dialects. The dataset is annotated with two different sets of guidelines: context-sensitive and context-insensitive. Preliminary results show that using all the resources together perform better on the context-sensitive dataset while the language models perform better on the context-insensitive dataset. Adding Sound-Change-Rules never hurts the performance yet their added value depends on how dialectal the dataset is since they only affect dialectal tokens. These results are encouraging given the different challenges that written Arabic impose. We plan on further extending our approach by identifying LCS on the sentence as well as the document level in addition to classifying the dialects.

[5]Tokens that were annotated as "MSA-DA" are counted twice, hence the *G-Tot.* count differs across the No-Context and Contextual annotations (Since the no-context annotation has more "MSA-DA" tokens. Also if a token has an actual class of MSA and the system produces "MSA-DA", it is considered a true-positive for MSA and false-positive for DA.

[6]This work is supported by the Defense Advanced Research Projects Agency (DARPA) BOLT program under contract number HR0011-12-C-0014.

References

Biadsy, F., Hirschberg, J., and Habash., N. (2009). Spoken arabic dialect identification using phonotactic modeling. In *Proceedings of the Workshop on Computational Approaches to Semitic Languages at the meeting of the European Association for Computational Linguistics (EACL), Athens, Greece.*

Chan, J. Y. C., Ching, P. C., LEE, T., and Meng, H. M. (2004). Detection of language boundary in code-switching utterances by bi-phone probabilities. In *Proceedings of the International Symposium on Chinese Spoken Language Processing*.

Dasigi, P. and Diab, M. (2011). Codact: Towards identifying orthographic variants in dialectal arabic. In *Proceedings of the 5th International Joint Conference on Natural Language Processing (ICJNLP), Chiangmai, Thailand*.

Diab, M., Hawwari, A., Elfardy, H., Dasigi, P., Al-Badrashiny, M., Eskandar, R., and Habash, N. (2013). Tharwa: A multi-dialectal multi-lingual machine readable dictionary. In *Forthcoming*.

Diab, M. and Kamboj, A. (2011). Feasibility of leveraging crowd sourcing for the creation of a large scale annotated resource for hindi english code switched data: A pilot annotation. In *Proceedings of the 9th Workshop on Asian Language Resources, Chiangmai,Thailand*.

Elfardy, H. and Diab, M. (2012). Simplified guidelines for the creation of large scale dialectal arabic annotations. In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC), Istanbul, Turkey.*

Ferguson (1959). *Diglossia. Word 15. 325340.*

Habash, N. (2007). *Arabic Morphological Representations for Machine Translation.*

Joshi, A. K. (1985). Processing of sentences with intrasentential code switching. In *R. Dowty, L. Karttunen, and A. M., Zwicky, eds., Natural Language Parsing: Psychological, Computational, and Theoretical Perspectives. Cambridge: Cambridge University Press. 190-205.*

Lyu, D.-C., yuan Lyu, R., chin Chiang, Y., and nan Hsu, C. (2006). Speech recognition on code-switching among the chinese dialects. In *Proceedings of the International Conference on nAcoustics, Speech and Signal Processing(ICASSP).*

Maamouri, M., Graff, D., Bouziri, B., Krouna, S., Bies, A., and Kulick, S. (2010). Ldc standard arabic morphological analyzer (sama) version 3.1.

Manandise, E. and Gdaniec, C. (2011). Morphology to the rescue redux: Resolving borrowings and code-mixing in machine translation. In *SFCM'11*, pages 86–97.

Parker, R., Graff, D., Chen, K., Kong, J., , and Maeda, K. (2009). Arabic gigaword fourth edition.

Rashwan, M., Al-Badrashiny, M., Attia, M., Abdou, S., and Rafea, A. (2011). A stochastic arabic diacritizer based on a hybrid of factorized and unfactorized textual features. In *Proceedings of the International Conference on Natural Language Processing and Knowledge Engineering*.

Solorio, T. and Liu, Y. (2008a). Learning to predict code-switching points. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), Honolulu, Hawaii*.

Solorio, T. and Liu, Y. (2008b). Part-of-speech tagging for english-spanish code-switched text. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), Honolulu, Hawaii*.

Stolcke, A. (2002). Srilm an extensible language modeling toolkit. In *Proceedings of the International Conference on Spoken Language Processing (ICSLP)*.