Colloquialising Modern Standard Arabic Text for Improved Speech Recognition

Sarah Al-Shareef,1,2, Thomas Hain1
1 Computer Science Department, The University of Sheffield, Sheffield, UK
2 Computer Science Department, Umm AlQura University, Makkah, Saudi Arabia
s.alshareef@gmail.com, t.hain@sheffield.ac.uk

Abstract
Modern standard Arabic (MSA) is the official language of spoken and written Arabic media. Colloquial Arabic (CA) is the set of spoken variants of modern Arabic that exist in the form of regional dialects. CA is used in informal and everyday conversations while MSA is formal communication. An Arabic speaker switches between the two variants according to the situation. Developing an automatic speech recognition system always requires a large collection of transcribed speech or text, and for CA dialects this is an issue. CA has limited textual resources because it exists only as a spoken language, without a standardised written form unlike MSA. This paper focuses on the data sparsity issue in CA textual resources and proposes a strategy to emulate a native speaker in colloquialising MSA to be used in CA text normalisation. The empirical results in Levantine CA show that using LMs estimated from colloquialised MSA data outperformed MSA LMs with a perplexity reduction up to 68% relative. In addition, interpolating colloquialised MSA LMs with a CA LMs improved speech recognition performance by 4% relative.

Index Terms: Colloquial Arabic, dialectical Arabic, language modelling, transfer learning, machine translation

1. Introduction
Modern standard Arabic (MSA) is the official language for 22 countries with around 300 million speakers. MSA is taught in schools and used for formal written and oral communication and discussions such as lectures, public speeches, news, magazines, and books. MSA is almost never the mother-tongue of speakers, but is only learnt at school. Colloquial Arabic (CA) is not one language, but is the set of spoken variants of modern Arabic that exist in the form of regional dialects and are considered generally to be mother-tongues in those regions. CA differs significantly from MSA phonetically, morphologically and syntactically. CA is also referred to interchangeably as dialectical Arabic and conversational Arabic as the variants can be considered as strong dialects and it is used mainly for conversations.

Native Arabic speakers can easily switch between two variants according to the situation and consequently can swap an utterance from one variant to the other. Transferring a given MSA utterance to a CA utterance is known as colloquialisation of MSA, while the reverse process is called CA normalisation.

Since CA exists only as a spoken language and without a standardised written form, it has limited textual resources. Furthermore, an MSA writing convention is imposed to represent the variants phonetically, which naturally has limitations on consistency and clarity. The only available resources that are also accessible to public research were collected from previous research effort in CA linguistic tools such as a set of transcribed telephone conversations, which sums into less than 2.5 million words with an average of 900k words for each dialect [2]. Such scarcity in textual data poses a challenge for developing an automatic speech recognition (ASR) for CA. This work focuses on the data sparsity issue and explores how statistical machine translation (SMT) framework can be employed in order to colloquialise MSA rich resources to be used in CA language models (LMs).

The rest of the paper is organised as follows. We start with a discussion of reported attempts to use existing MSA textual data for developing a language model for CA (§2). How pairs of CA and MSA sentences were collected with crowdsourcing framework in order to be used as training samples for the colloquialisation model will be described in §3. This is followed by a description of the proposed strategy to narrow the gap between MSA and CA through colloquialisation within a SMT framework in order to generate more CA textual data (§4). Language models based on colloquialised MSA resources were empirically evaluated and the results are reported in §5 and §6. The paper is concluded in §7.

2. Related Research

Given that CA resources are limited and MSA resources are plentiful several studies explored their use to enrich CA data for developing natural language processing tools. Much work explored using MSA data either directly, by finding a mapping between CA and MSA, or by parsing CA and MSA and using the syntactic and morphological level instead of or with the lexical level.

Just pooling of transcribed CA text with MSA data, for example Egyptian CA (ECA) with MSA [3] and Qatari CA [4], yielded an insignificant (if any) reduction in LM perplexity. Similar outcomes were observed when Kirchhoff et al. [3] interpolated two LMs, one estimated from a small ECA training set and the other estimated from MSA data, with optimised weights even when the chosen MSA data were selected to be conversational in nature.

Other studies in the context of MT attempted to transform CA into MSA due to the absence of CA-English parallel corpora. Motivated by the rich MSA-English MT resources, many researchers transformed CA to MSA, i.e. CA normalisation, in order to be able to use existing MSA resources. For instance, [5] employed a hybrid normalisation approach to normalise ECA, which applied a combination of mapping rules and a statistical tokenising and tagging model trained on an ECA morphological lexicon. Another hybrid normalisation approach was proposed by [6]. Here the normalisation method transferred CA words to...
MSA based on character- and morpheme-level mapping rules. Afterwards an MT system was used to translate from MSA to English. While [6] normalised both affixes and stems to MSA vocabulary, [7] only applied mapping rules on the affixes but also used morphological analysis information and dictionaries in addition to language models and allowed multiple morphological analyses in the form of lattices to be translated by an MT system to English.

With the emergence of social media, more written CA can be observed where users use their own mother-tongue, namely CA, in conversational responses. [8] harvested the web for ECA and MSA lexicons, while the COLABA project [9] constructed similar resources from web logs. Based on their experience, [10] composed a set of guidelines for constructing such resources with the aid of automatic dialect identifiers.

3. Constructing a CA-MSA parallel corpus

Unlike MSA, CA lacks standard conventions for writing colloquial words. Therefore, native Arabic writers usually improvise the spelling of such words and this leads to noisy and unreliable colloquialised MSA texts. Hence, for creating a parallel CA-MSA corpus, colloquialisation of MSA data is more problematic than normalisation of CA data for the consistency of annotation conventions, and the latter is used here – with the assumption that the process is somewhat invertible.

Lately, crowdsourcing platforms, such as Amazon’s Mechanical Turk (AMTurk), are used for collecting and annotating resources for computational linguistics (e.g. [11, 12]). [13] and [14] provided general guidelines for best practice in using such platforms in order to obtain high quality NLP resources. Crowdsourcing allows annotation tasks to be distributed among several non-professional annotators by splitting them into smaller tasks, known as microtasks. Unfortunately AMTurk is restricted for use by USA residents only; therefore, the Upwork1 platform was employed instead. Upwork is an international work platform to connect freelancers and work contractors together. Unlike AMTurk, Upwork does not scale easily to large numbers of annotators because each of them needs an individual contract before enrolling and performing any task. Nevertheless, the experience level of hired annotators, hence their outcome quality, is much higher in Upwork than in AMTurk. Moreover, the cost of performing the normalisation of CA using Upwork remains considerably lower than hiring professionals.

3.1. Data selection

Normalisation of CA requires sentences in CA which were drawn two Levantine CA (LCA) corpora, Fisher2 and Appen3. Both corpora are distributed by the LDC. The data represents conversations by native LCA speakers talking to their friends and families, as well as unrelated individuals, about topics suggested by the corpus collectors. The two sets were merged into one set as trainLCA (Table 1). A subset of these transcriptions was selected to avoid repetitions and to ensure more lexical coverage. Since CA and MSA share more than 60% of their vocabulary [15], only sentences with at least one non-MSA word are included in the chosen set. A background MSA lexicon of 2.5M words was constructed from MSA resources4. A word was considered an MSA word if it was found in the MSA lexicon; otherwise, it is assumed to be a CA word. Usually, sentences in CTS

---

1 wwww.upwork.com
2LDC2006S29, LDC2006T07
3LDC2007S01 & LDC2007T01

---

Figure 1: Task design for collecting parallel CA-MSA text using crowdsourcing.
First, for each \( c \), a modified average edit-distance (MAED) is computed between each pair of its normalisations, for instance, \( s_i \) and \( s_j \), then the similarity between them is computed by MAED\((s_i, s_j)\) as follows:

\[
\text{MAED}(s_i, s_j) = 1 - \frac{M - \text{ED}(s_i, s_j)}{0.5(N_i + N_j)},
\]

where \( \text{ED}(s_i, s_j) \) is the total number of edits, including insertions, deletions and substitutions, between \( s_i \) and \( s_j \), \( M \) is the number of matched words between \( s_i \) and \( s_j \) and \( N_i \) and \( N_j \) is the number of words in \( s_i \) and \( s_j \), respectively. The two normalisations are accepted and \( n \) is removed from the normalisation pool if \( s_i \) and \( s_j \) are similar, i.e. MAED\((s_i, s_j)\) < \( \alpha \) where \( \alpha \) were chosen empirically as 0.3. If not, the decision was made based on the quality of annotation which was measured by the annotator’s normalisation for the GS sentences. Annotator’s quality in the job, \( Q(x) \) is computed as follows:

\[
Q(x) = \sum_{c \in C_x} |c|^{-1} \sum_{T \in c} \text{MAED}(T, t_x),
\]

where \( \text{MAED}(T, t_x) \) is computed using Eq.1 between an annotator’s normalisation \( t_x \) for \( c \) and its GS normalisations \( T \). \( |c| \) is the number of GS normalisations \( T \) for \( c \). This quantity is then averaged by the number of \( c \) in the job, \( \sum_{C_x} |c| \), typically 3. The annotator normalisations were accepted if \( Q(x) < 0.5 \); otherwise, the sentence \( n \) is added to the normalisation pool to be considered for normalisation again.

For the validation task, a CA sentence with all its normalisations, \((n, s_i, s_j, ...)\), was provided. An annotator can reject or accept each \( s_i \) as a valid normalisation. If no normalisations survived during the process, \( n \) is returned to the normalisation pool again.

3.3. Data
Using the Upwork platform, 47 native LCA speakers were enrolled to normalise a set of 20379 sentences with a total of 142318 words. The normalised LCA set has 147007 words with an average of 1.4 normalisations per LCA sentence.

4. Colloquisation system
The outcome of the previous process was parallel corpus of LCA-MSA data, which is a set of pairs of LCA sentences along with its normalised variants. A translation model was estimated as a colloquisation model based on the crowdsourced parallel corpus using an SMT framework, based on the Moses toolkit [16]. The source language was MSA (i.e. normalised variant) and the target language was LCA. The colloquisation model obtained a BLEU score of 0.994 on testLCA (Table 4).

5. Colloquised MSA language model
After estimating the colloquisation model, two MSA resources, NW10 and BC, were colloquialised with the model and the Moses decoder. BC is subset of GALE Arabic Broadcast Conversations and NW10 is a subset of Arabic Gigaword

Table 1: Data sets used for training and testing.

<table>
<thead>
<tr>
<th></th>
<th>trainLCA</th>
<th>BC</th>
<th>NW10</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sentences</td>
<td>433,076</td>
<td>89,816</td>
<td>1,477,544</td>
</tr>
<tr>
<td>Words</td>
<td>1,906,286</td>
<td>1,433,932</td>
<td>15,779,447</td>
</tr>
<tr>
<td>Vocabulary</td>
<td>81,636</td>
<td>102,629</td>
<td>424,922</td>
</tr>
<tr>
<td>word/sentence</td>
<td>4.4</td>
<td>16.0</td>
<td>10.7</td>
</tr>
</tbody>
</table>

Figure 2: Schematic diagram for developing a language model based on colloquialised MSA text. The “+” sign indicate a linear interpolation between two LMs.

newswire resources; both sets are described in Table 1. The resulting colloquialised MSA corpora were employed to estimate a standard trigram LM using modified Kneser-Ney discounting and backoff for each corpus. These LMs were estimated using SRILM toolkit [17]. Figure 2 illustrates the development process for the colloquialised MSA-based LM. A vocabulary of 41688 words was chosen by keeping all non-singletons from trainLCA. All LMs were estimated using the same vocabulary, which has an OOV rate of 2.5% on the testLCA (Table 4). Our baseline LCA LM, which was estimated from trainLCA data only without interpolation with any MSA resources, has a perplexity of 213.3 on testLCA.

Table 2a shows the perplexity computed over different interpolation configurations of BC, NW10 and LCA LMs on testLCA. Although BC is an MSA resource, its perplexity is equivalent to almost one tenth of that of NW10. This is mainly because the style of the BC dataset is conversational while NW10 is intended for written media and thus has a much richer context than that of BC. Consequently, the BC LM was assigned a higher interpolation weight than the NW10 LM when they were linearly interpolated with the LCA LM. The interpolated LM gave a relative perplexity improvement of 1.9% and 3.0% respectively and 3.4% when both were included in the interpolation to reach a perplexity of only 206. LMs estimated from colloquialised MSA resources showed a considerable reduction in the perplexity, especially for the NW10 dataset, as shown in Table 2b. In comparison to the perplexity computed from BC and NW10 (shown in Table 2a), a relative reduction of 30% and 68% respectively was obtained with colloquialised corpora instead of equivalent MSA data. Moreover, the obtained reduction in the perplexity resulting from interpolating the LCA LM and LMs estimated from colloquialised MSA resources was twice that of an interpolation with LMs estimated from MSA resources directly. Table 3 lists the relative difference in the number of \( n \)-grams of order 1 to 3 found in LMs estimated from MSA text and LMs estimated from colloquialised MSA text. As shown in the table, the number of both bigrams and trigrams were increased by at least 1.7% and 3.6% respectively depending on the size of the colloquialised dataset. This empirically proved that automatically colloquialised MSA text can be used as an additional resource for developing CA LMs.

6. Speech recognition experiments
The data used for training acoustic models was drawn from Fisher LCA corpus, which consists of 143.3 hours of conversational telephone speech (CTS) recordings. As described in a previous work on Fisher corpus [18], a test set of 5.1 hours was
Table 2: Perplexity and recognition results when using an interpolated LCA LM with different combinations of LM components estimated on (a) MSA resources or (b) colloquialised MSA resources. If the interpolation weight is 1.0 that means there is no interpolation with any other component.

(a) MSA resources

<table>
<thead>
<tr>
<th>Interpolation weights</th>
<th>LCA</th>
<th>BC</th>
<th>NW10</th>
<th>Perplexity</th>
<th>PLP</th>
<th>PLP+BN</th>
<th>WER</th>
<th>PLP</th>
<th>PLP+BN</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.0</td>
<td>213.3</td>
<td>60.3</td>
<td>54.4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1.0</td>
<td>2066.8</td>
<td>gray?75</td>
<td>gray?75</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1.0</td>
<td>19474.4</td>
<td>gray?75</td>
<td>gray?75</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.955</td>
<td>0.005</td>
<td>209.2</td>
<td>59.8</td>
<td>54.0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.962</td>
<td>0.038</td>
<td>206.9</td>
<td>60.2</td>
<td>54.3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.947</td>
<td>0.028</td>
<td>0.026</td>
<td>206.0</td>
<td>59.7</td>
<td>54.0</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

(b) colloquialised MSA resources

<table>
<thead>
<tr>
<th>Interpolation weights</th>
<th>LCA</th>
<th>BC</th>
<th>NW10</th>
<th>Perplexity</th>
<th>PLP</th>
<th>PLP+BN</th>
<th>WER</th>
<th>PLP</th>
<th>PLP+BN</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.0</td>
<td>213.3</td>
<td>60.3</td>
<td>54.4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1.0</td>
<td>1452.6</td>
<td>gray?75</td>
<td>gray?75</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1.0</td>
<td>6304.7</td>
<td>gray?75</td>
<td>gray?75</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.932</td>
<td>0.068</td>
<td>206.5</td>
<td>58.4</td>
<td>52.8</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.933</td>
<td>0.067</td>
<td>200.4</td>
<td>59.9</td>
<td>54.2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.915</td>
<td>0.033</td>
<td>0.052</td>
<td>199.5</td>
<td>57.1</td>
<td>52.8</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 3: Relative difference in the number of n-grams (of order 1 to 3) between LMs estimated from MSA resources (baseline) and LMs estimated from colloquialised MSA resources.

<table>
<thead>
<tr>
<th>Corpus</th>
<th>unigrams</th>
<th>bigrams</th>
<th>trigrams</th>
</tr>
</thead>
<tbody>
<tr>
<td>BC</td>
<td>0.0</td>
<td>+1.7%</td>
<td>+3.6%</td>
</tr>
<tr>
<td>NW10</td>
<td>0.0</td>
<td>+6.1%</td>
<td>+4.9%</td>
</tr>
</tbody>
</table>

Table 4: Training and testing data sets for ASR experiments.

<table>
<thead>
<tr>
<th>Words</th>
<th>Vocabulary</th>
<th>Hours</th>
</tr>
</thead>
<tbody>
<tr>
<td>FisherLCA</td>
<td>1528342</td>
<td>67195</td>
</tr>
<tr>
<td>testLCA</td>
<td>53644</td>
<td>8762</td>
</tr>
</tbody>
</table>

Acknowledgments

The first author would like to thank Umm AlQura University, Makkah, Saudi Arabia for funding this work as part of her PhD scholarship programme.
9. References


