INTERSPEECH 2016
September 8–12, 2016, San Francisco, USA

Colloquialising Modern Standard Arabic Text for Improved Speech Recognition

Sarah Al-Shareef[1,2], *Thomas Hain*[1]

[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 language models (LMs) by use of a machine translation (MT) framework. 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 [1]. 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 normally 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 dialects [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

Copyright © 2016 ISCA

http://dx.doi.org/10.21437/Interspeech.2016-788

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 mircotasks. Unfortunately AMTurk is restricted for use by USA residents only; therefore, the Upwork[1] 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, Fisher[2] and Appen[3]. 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 insure 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 resources[4]. 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 corpora are short in length (4-6 words per sentence) and in undiacritised form. This imposes a challenge for the annotator to choose the corresponding MSA match that serves the intended meaning. Therefore, the one preceding and one subsequent sentence were presented as well, to provide some semantic context. Because there might be more than one valid normalisation for a single CA sentence, more than one normalisation was allowed to gain a much richer mapping between CA and MSA.

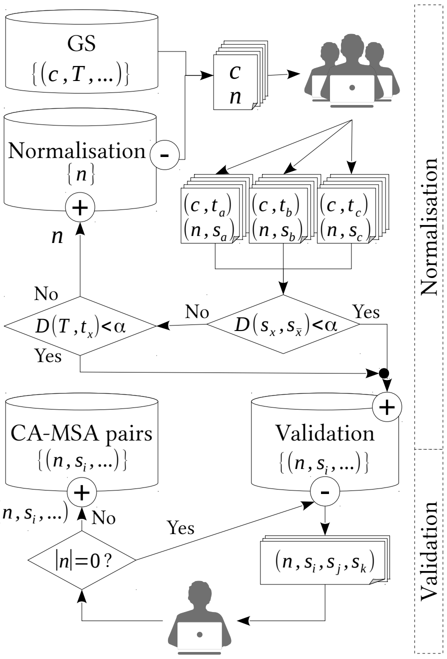

Figure 1: *Task design for collecting parallel CA-MSA text using crowdsourcing.*

3.2. Quality control and task design

Several quality measures were used to ensure annotation quality. First, enrolled annotators had to be native speakers of the presented CA dialect. Second, following the guidelines of [13], several control sentences with gold standard (GS) MSA equivalences were hidden in the job's set - each job has 10 sentences with at least 3 GS sentences. An additional quality control procedure was to provide the source CA sentence and all its normalised variants resulting from a previous normalisation task so that any invalid normalisation variants would be rejected. If none of the normalisation variants survived, that CA sentence was returned to be normalised again.

Figure 1 illustrates how each task was constructed and the interaction with the crowd. From the selected subset, 2000 sentences were randomly selected, rendered into MSA manually and kept in a GS sentence pool. Each GS sentence is shown as c and its normalisations are T. Apart from GS sentences, all sentences, n, were initially assigned to the normalisation pool. Each n was assigned to at least 3 annotators.

In the normalisation task, annotators were asked to render each colloquial word into its MSA equivalent using an undiacritised form such that one LCA word can be rendered to a phrase of more than one word in MSA and vice versa. In addition, annotators were not encouraged to reorder the normalised phrase unless it was completely unacceptable in MSA, which is rare given the syntactic flexibility of MSA.

For each n and c, each annotator provided a normalisation s_x and t_x, respectively, where x denoted the annotator ID.

[1] *www.upwork.com*

[2] LDC2006S29, LDC2006T07

[3] LDC2007S01 & LDC2007T01

[4] Arabic Gigaword: LDC2011T11 & BAMA:LDC2010L01.

Table 1: *Data sets used for training and testing.*

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

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 $\mathrm{MAED}(s_i, s_j)$ as follows:

$$\mathrm{MAED}(s_i, s_j) = 1 - \frac{M - \mathrm{ED}(s_i, s_j)}{0.5(N_i + N_j)}, \quad (1)$$

where $\mathrm{ED}(s_i, s_j)$ 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. $\mathrm{MAED}(s_i, s_j) < \alpha$ where α 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:

$$Q(x) = |C_x|^{-1} \sum_{c \in C_x} |c|^{-1} \sum_{T \in c} \mathrm{MAED}(T, t_x), \quad (2)$$

where $\mathrm{MAED}(T, t_x)$ 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, $|C_x|$, 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_x 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. Colloquialisation system

The outcome of the previous proces 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 colloquialisation 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 colloquialisation model obtained a BLEU score of 0.994 on testLCA (Table 4).

5. Colloquialised MSA language model

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

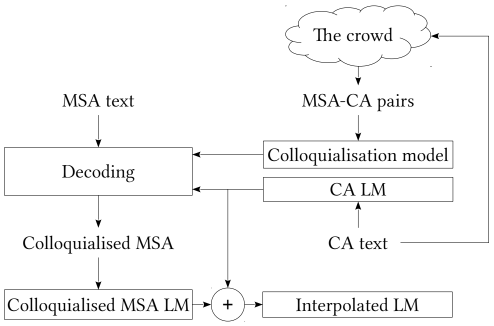

Figure 2: *Schematic diagram for developing a language model based on colloquialised MSA text. The "+" sign indicate a linear interpolation between two LMs.*

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

[5] LDC2013T04

Table 2: *Perplexity and recognition results when usig 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*

Interpolation weights				WER	
LCA	BC	NW10	Perplexity	PLP	PLP+BN
1.0			213.3	60.3	54.4
	1.0		2066.8	gray!75	gray!75
		1.0	19474.4	gray!75	gray!75
0.955	0.005		209.2	59.8	**54.0**
0.962		0.038	206.9	60.2	54.3
0.947	0.028	0.026	206.0	**59.7**	**54.0**

(b) *colloquialised MSA resources*

Interpolation weights				WER	
LCA	BC	NW10	Perplexity	PLP	PLP+BN
1.0			213.3	60.3	54.4
	1.0		1452.6	gray!75	gray!75
		1.0	6304.7	gray!75	gray!75
0.932	0.068		206.5	58.4	**52.8**
0.933		0.067	200.4	59.9	54.2
0.915	0.033	0.052	199.5	**57.1**	**52.8**

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.*

Corpus	unigrams	bigrams	trigrams
BC	0.0	+1.7%	+3.6%
NW10	0.0	+6.1%	+4.9%

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

	Words	Vocabulary	Hours
FisherLCA	1528342	67195	143.3
testLCA	53644	8762	5.9

constructed by random selection of conversation sides to maintain a homogeneous and balanced recording conditions. Both sets are described in Table 4. A pronunciation dictionary was constructed by converting the automatically diacritised transcriptions, using CRF-based diacritiser [19], into phonemes using a set of pronunciation rules and forced-alignment to overcome all silent and ambiguous graphemes. The recognition dictionary has an average of 2.5 pronunciation per word using 39 phonemes. As aforementioned, all LMs were estimated using the same vocabulary, which has an OOV rate of 2.5%.

The audio data was segmented using timings available in the corpus and two types of acoustic features were extracted. First, 13 PLP features plus their 1st and 2nd derivatives; second, the PLP features were concatenated with 26 bottleneck features (PLP+BN), which were extracted from a 4-hidden layer DNN. 31 adjacent frames (15 frames to the left and 15 frames to the right) of 23 dimensional log Mel filter bank features were concatenated to form a 713dimensional super vector; DCT was applied to this super vector to decorrelate and compress it to 368 dimensions and then fed into the neural network. The network was trained on 3800 triphone state targets and the 26 dimensional bottleneck layer was placed before the output layer. The objective function used was framelevel crossentropy and the optimisation was done with stochastic gradient descent and the backpropagation algorithm. DNN training was performed with the TNet toolkit [20]. Cepstral mean and variance normalisation was applied on a per conversation side basis for the PLP system only. All models are trained using a standard mixup maximum likelihood regime with left and right context trigraphemes using HTK toolkit [21]. Left-to-right HMMs with 3 emitting states were used and clustered at the state level using a binary decision tree with phonologically motivated questions. After a gradual increase of Gaussian mixture components models contained 16-mixture components for each state with 3800 clustered states.

The recognition results are shown in the last columns of Table 2a when MSA resource were used directly and Table 2b when MSA resources were colloquialised. The maximum improvement gained from using MSA components over LCA LM alone was 1% WER relative. Using colloquialised MSA resources were found to be beneficial for both systems but with different degrees. For instance, the improvement in recognition performance increased from 0.8-0.9% WER relative to reach 3.1-3.5% WER relative when both MSA resources were colloquialised and interpolated with LCA LM.

Although NW10 LM was estimated from a volume of text that is 10 folds more than that of BC LM, it did not contribute in reducing neither the perplexity or WER. Even colloquialisation did not help in improving recognition performance of NW10 LM significantly. This can be accounted for conversational nature of BC data which matched those telephone conversations the colloquialisation model was estimated from. Evidently, colloquialising BC data narrow the gap between MSA and LCA which is reflected by the improvement gained in both perplexity and WER.

7. Discussion and conclusion

The main objective of this paper was to investigate the exploitation of the large volume of text in a rich-resourced language, MSA, for developing a LM for an under-resourced variant of the language, CA, to improve both LM and ASR performance. The proposed strategy generated additional in-domain data which was used as training materials for estimating LMs, and improved word estimates without expanding the lexical coverage. The resulting LMs outperformed LMs estimated from the MSA data with a perplexity reduction up to 68% relative. Moreover, the perplexity reduction obtained from interpolating these colloquialised MSA LMs was twice that obtained from interpolating MSA LMs to reach 6.5% relative in comparison to the baseline CA LM. The perplexity reduction improved the ASR performance in LCA CTS task with 3.4% WER relative.

The impact of the proposed colloquialisation on ASR performance was more prominent when the MSA source was in conversational style, i.e. using 1^{st} and 2^{nd} person sounds. On the other hand, when the MSA source was mainly in 3^{rd} person sound such as in broadcast news and newswire, the impact was far less because the colloquialisation model was not trained on such data. The proposed colloquialisation model learned the mapping between the MSA and CA conversations only. If the same strategy is applied to transfer the sound and style of the data, much richer resources can be exploited in developing more appropriate interactive interfaces.

8. 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

[1] Janet C. E. Watson, *The phonology and morphology of Arabic*, Oxford University Press, New York, 2002.

[2] Nizar Habash, *Introduction to Arabic Natural Language Processing*, Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers, 2010.

[3] K Kirchhoff, J Bilmes, S Das, N Duta, M Egan, G Ji, F He, J Henderson, D Liu, and M Noamany, "Novel approaches to Arabic speech recognition: report from the 2002 Johns-Hopkins summer workshop," in *Proc ICASSP*, 2003, vol. 1, pp. 344–347.

[4] Mohamed Elmahdy, Mark Hasegawa-Johnson, and Eiman Mustafawi, "A transfer learning approach for under-resourced Arabic dialects speech recognition," in *Workshop on Less Resourced Languages, new technologies, new challenges and opportunities (LTC 2013)*, 2013.

[5] Hitham Abo Bakr, Khaled Shaalan, and Ibrahim Ziedan, "A hybrid approach for converting written Egyptian colloquial dialect into diacritized Arabic," in *Proc 6th Int Conf Informatics and Systems*, Cairo, Egpyt, 2008, pp. 27–33.

[6] Hassan Sawaf, "Arabic dialect handling in hybrid machine translation," in *Proc Conf Assoc Machine Translation in the Americas (AMTA)*, Denver, CO, 2010.

[7] Wael Salloum and Nizar Habash, "Dialectal to standard Arabic paraphrasing to improve Arabic-English statistical machine translation," in *the First Workshop on Algorithms and Resources for Modelling of Dialects and Language Varieties*, Edinburgh, Scotland, 2011, pp. 10–21.

[8] Rania Al-Sabbagh and Roxana Girju, "Mining the web for the induction of a dialectical Arabic lexicon," in *Proc LREC*, Valletta, Malta, 2010, pp. 288–293.

[9] Mona Diab, Nizar Habash, Owen Rambow, Mohamed Altantawy, and Yassine Benajiba, "COLABA: Arabic dialect annotation and processing," in *Proc LREC Workshop on Semitic Language Processing*, 2010, pp. 66–74.

[10] Heba Elfardy and Mona T Diab, "Simplified guidelines for the creation of large scale dialectal Arabic annotations.," in *Proc LREC*, 2012, pp. 371–378.

[11] Alexander Sorokin and David Forsyth, "Utility data annotation with Amazon Mechanical Turk," *Urbana*, vol. 51, no. 61, pp. 820, 2008.

[12] Gaurav Kumar, Yuan Cao, Ryan Cotterell, Chris Callison-Burch, Daniel Povey, and Sanjeev Khudanpur, "Translations of the CALLHOME Egyptian Arabic corpus for conversational speech translation," in *IWSLT*, 2014.

[13] Omar F Zaidan and Chris Callison-Burch, "Crowdsourcing translation: Professional quality from non-professionals," in *Proc 49th Annual Meeting Assoc Computational Linguistics*. Association for Computational Linguistics, 2011, vol. 1 - Human language technologies, pp. 1220–1229.

[14] Marta Sabou, Kalina Bontcheva, Leon Derczynski, and Arno Scharl, "Corpus annotation through crowdsourcing: Towards best practice guidelines," in *Proc LREC*, 2014, pp. 859–866.

[15] Nizar Habash and Owen Rambow, "MAGEAD: a morphological analyzer and generator for the Arabic dialects," *Proc ACL*, Jul 2006.

[16] Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, et al., "Moses: Open source toolkit for statistical machine translation," in *Proc ACL*. Association for Computational Linguistics, 2007, pp. 177–180.

[17] Andreas Stolcke, "SRILM – an extensible language modeling toolkit," in *Proc INTERSPEECH*, 2002, pp. 901–904.

[18] S. Al-Shareef and T. Hain, "An Investigation in Speech Recognition for Colloquial Arabic," in *INTERSPEECH*, 2011, pp. 2869–2872.

[19] S. Al-Shareef and T. Hain, "CRF-based Diacritisation of Colloquial Arabic for Automatic Speech Recognition," in *INTERSPEECH*, 2012, pp. 1824–1827.

[20] Karel Vesel, Luks Burget, and Frantisek Grzl, "Parallel training of neural networks for speech recognition.," in *INTERSPEECH*. 2010, pp. 2934–2937, ISCA.

[21] S. J. Young, G. Evermann, M. J. F. Gales, T. Hain, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. C. Woodland, *The HTK Book, version 3.4*, Cambridge University Engineering Department, Cambridge, UK, 2006.