Improving Egyptian-to-English SMT by mapping Egyptian into MSA

Nadir Durrani* — University of Edinburgh — dnadir@inf.ed.ac.uk

Yaser Al-Onaizan, Abraham Ittycheriah — IBM T.J. Watson Research Center — {onaizan,abei}@ibm.com

Abstract. One of the aims of DARPA BOLT project is to translate the Egyptian blog data into English. While the parallel data for MSA[1]-English is abundantly available, sparsely exists for Egyptian-English and Egyptian-MSA. A notable drop in the translation quality is observed when translating Egyptian to English in comparison with translating from MSA to English. One of the reasons for this drop is the high OOV rate, where as another is the dialectal differences between training and test data. This work is focused on improving Egyptian-to-English translation by bridging the gap between Egyptian and MSA. First we try to reduce the OOV rate by proposing MSA candidates for the unknown Egyptian words through different methods such as spelling correction, suggesting synonyms based on context etc. Secondly we apply convolution model using English as a pivot to map Egyptian words into MSA. We then evaluate our edits by running decoder built on MSA-to-English data. Our spelling-based correction shows an improvement of 1.7 BLEU points over the baseline system, that translates unedited Egyptian into English.

1 Introduction

The use of dialectal Arabic has been previously only limited to the speech whereas written texts were produced using MSA. With the rapidly increasing availability of the colloquial text, due to influx of social media in the Arabic-speaking countries in recent times, there has been interest in translating forums and blogs. DARPA GALE project aimed at translating news wire and parliamentary proceedings. The focus in BOLT project has shifted towards translating blog data and different dialects of Arabic, more specifically Egyptian which is considered to be the most widely used dialect after MSA. The new focus of translating dialect and blog data presents numerous challenges. An immediate bottle-neck is the lack of NLP resources for various dialects. Secondly the blog data is user-generated therefore noisy and lack standardization in orthography [1].

While the parallel data for MSA-English is abundantly available, sparsely exists for Egyptian and other dialects of Arabic. A notable drop in the quality of translation is observed when translating Egyptian blog data using translation models built on top of MSA-English parallel data. Table 1 shows results, in terms of BLEU [2], when decoding MSA (Gale-dev10 set) and Egyptian (tahyyes dev set) using two different decoders

* Work done during an internship at IBM T.J. Watson Research Center

[1] MSA = Modern Standard Arabic

TRL [3] and DTM [4]. Some of this drop can be attributed to the OOV rate which is high as 3.66% when translating Egyptian. While the rest occurs because of the dialectal differences and data mismatch

The comparison of BLEU scores in Table 1, however, may not be fair because it is obtained from running decoders on different dev sets. In order to ensure that these numbers reflect a true comparison, we did a pilot study. We randomly took a set of 100 Egyptian sentences and got them translated to MSA through a human translator. We then decoded both MSA and EGY sentences into English which used models trained on MSA-En parallel data. Both TER [5] and BLEU scores (Table 2) for the Egyptian sample were worse by more than 3 points as compared to that of the output of MSA. The high OOV rate (3.68%) in the English output of Egyptian sample, also confirms the result in Table 1.

Table 1. Baseline Comparison for MSA and EGY (Egyptian)

	TRL	**DTM**	**OOV Rate**
MSA	22.83	28.81	0.64%
EGY	19.49	22.29	3.66%

Table 2. Translating 100 Sentences of MSA and EGY into English

TRL	**TER**	**BLEU**	**OOV Rate**
MSA	58.52	18.77	1.07%
EGY	61.75	15.22	3.68%

Our focus in this paper is to improve Egyptian-to-English translation by bridging the gap between Egyptian and MSA. We try to learn dialectal differences between Egyptian and MSA and map the Egyptian words, that our system does not know how to translate or translate well, to their corresponding MSA words.

The paper is organized as follows. In Section 2 we provide a review of previous work In Section 3, we study the patterns of OOV words in the Egyptian data, dialectal differences between Egyptian and MSA, and discuss methods to propose MSA candidates for these. In Section 4, we describe a model to rank the candidates in a stack-based decoding framework. In Section 5, we present results on the OOV handling. In Section 6 we try to exploit a very small Egyptian-English corpus, using English as a pivot to map Egyptian to MSA using the well-known convolution model . In Section 7 we conclude the paper.

2 Previous Work

A plentiful amount of research has been spent in developing resources and natural language processing of MSA [6]. However, research on different dialects of Arabic is relatively sparse [7, 8]. In this section we discuss previous work on machine translation of dialects. A hybrid machine translation that uses both rule-based and statistical methods to transform an Egyptian sentence into a diacritized MSA sentence was proposed in [9].

The input sentence is first tokenized and pos-tagged through a statistical model. A rule-based model, built on top of Egyptian-MSA lexicon, is then used to transfer the source into diacritized MSA. [10] also improved dialectal translation in hybrid machine translation by normalizing dialectal Arabic on character and morpheme level using a dialect-specific morphological analyzer. By applying their processing to the training and test corpora, they observed an improvement in the translation quality by approximately 2% on web text in terms of BLEU score. Like this paper, [11] focus on translating OOV words in the dialect Arabic. They propose paraphrases of the source language words. The candidates are obtained by applying morphological analysis on the input and mapping the affixes of OOV words into their MSA counterparts. The transformation is only applied to the affixes and not to the stems. The resulting candidates are then fed into MSA-to-English SMT as an input lattice. Their methodology gives an improvement of 0.56 BLEU points on a test set having an OOV Rate of 1.51%. In a recent effort, [1] built a 1.5 M words Dialectal (Levantine and Egyptian)-to-English data using Amazon's Mechanical Turk crowdsourcing service. In their study, they showed that a system built on small amount of dialectal data (1.5 Million words) improves the translation quality by more than 6 BLEU points than their MSA counterpart built from an enormous amount of data (150 Million Words). One of the interesting finds in their paper is that adding Dialect-English parallel data to the training (MSA-English parallel data) proves much more benificial than using Dialect-MSA parallel data to first transform dialect into MSA and then translating from MSA to English. The former is better than latter by a difference of roughly 2 BLEU points. [12] used character level transformation including morphological, phonological and spelling changes to narrow down the gap between Egyptian and MSA, subsequently improving Egyptian-to-English machine translation. Their work is similar to ours and goes in the direction of translating between the related languages using word-level translation, character level transformations, and language specific rules [13, 14].

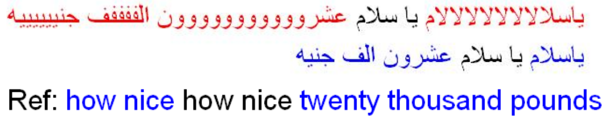
Other notable contributions towards building NLP resources for dialect discussed as follows. [7] built parser for spoken dialect using MSA tree-bank. [15] minned the web to extract a Dialect-to-MSA lexicon. A statistical morphological segmenter for Iraqi and Levantine speech transcripts was built by [16]. Their supervised algorithm for morpheme segmentation reduces the OOV rate.

3 Patterns of OOV and Methods for Candidate Suggestions

In this section, we study the patterns of errors in the Egyptian dev set and propose methods to give MSA candidates for these. After analyzing some data, we classified the error patterns into five sets i) substring repetition, ii) compounding errors iii) spelling differences, iv) dialect specific errors and v) true OOV words. We briefly discuss each of these categories.

3.1 Substring Repetition

Most noticeable but rather small number of errors in the blog data appeared due repetition of a character/substring in a known word. See Figure 1 for reference. In the first word, the susbstring لا repeats, in third, fourth and fifth words characters و, ف and ي[2] repeat. For a third consecutive appearance of a character or substring in a string, one can be sure that it is spurious and can be safely deleted. For all such instances of OOV words we remove the spurious characters one at a time and hypothesize the ones found in the MSA vocabulary.



ياسلالالالالالام يا سلام عشروووووووووووون الففففف جنيييييه

ياسلام يا سلام عشرون الف جنيه

Ref: how nice how nice twenty thousand pounds

Fig. 1. Character and Substring Repetition

3.2 Compounding Errors

A small number of errors were caused due to multiple words conjoined into a single token. For example in an OOV word بالحشدوالهتافات, بالحشد (crowd) and الهتافات (chants) are compounded into a single token through a joining morpheme و (and). We propose MSA candidates for such OOV words by splitting them into their right components following [17].

3.3 Spelling Differences

From the error analysis of several hundred Egyptian sentences we noticed some dialectal specific corrections that can be applied to transform an Egyptian word into its MSA counterpart. For example in the Egyptian word بالصهاينه (Zionist), character ب can be dropped from the beginning and the ending character ه can be changed to character ة to form an MSA word الصهاينة (Zionist). After looking at a sample of hundered MSA-Egyptian word pairs, a list of rules (given in Table 3) is extracted. We apply these transformation rules to the error word and hypothesize the ones found in the MSA vocabulary.

To do some further analysis, we extracted a list of 5000 most frequent Egyptian words, with context, from a 2 Million word monolingual corpus. We then got these translated into MSA through human translator. Of the 5000 words approximately 70% were translated to themselves (source word). Of the remaining 30% words that the human choose to translate differently, 33% can be transformed to MSA by applying single-edit distance to the original word. Another 16.4% can be converted to MSA by applying two edits. This provides a strong motivation to use the spelling correction mechanism

[2] The shapes of Arabic characters ف and ي change to ﻓ and ﻳ respectively, in context because of the cursive nature of Arabic script

Table 3. Dialectal Rules to Convert Egyptian to MSA

Egyptian	MSA
bXXX	XXX
XXXh	XXXp
XXXp	XXXh
XXXy	XXX +y
XXXNa	XXX +Na
H |h (y|n|t|A) XXX	s (y|n|t|A) XXX

as one of the techniques to propose MSA candidates. In our spelling correction module we apply all possible single edits (deletions, substitution and insertions) to the unknown word to get the candidate strings and hypothesize the ones that are found in the MSA vocabulary.

3.4 Dialect-based Errors and True OOVs

The fourth class of OOV words contain purely dialectal Egyptian words to which applying spelling correction does not yield an MSA word. For example Egyptian word مستنية (waiting) has an alternative منتظرة in MSA. Finally a portion of errors constitute name entities like سكسكة (Sekska). We call this fifth category of unknown words as True OOVs.

Context-based Synonym Suggestions In order to specifically target the last two classes of errors we use a technique similar to proximity based synonym acquisition [18, 19]. The idea is to propose a synonym for an unknown word based on the language model context. The intuition is that synonyms of a word are likely to share the same context. For example consider a sentence “Barking like a tyke”. Assume that “tyke” is unknown to our translation model. Based on the context “Barking like a” we might be able to produce candidates like “dog”, “doggy”, “bitch” etc. We do not use a fixed radius of words to propose candidates but also take advantage of language model back off with smoothing when proposing candidates.

4 Model

4.1 Egyptian-to-MSA Decoder

In the last section we discussed a bunch of methods to propose candidates for the unknown Egyptian words. Now we devise a model to score these candidates. Say we observe an Egyptian sentence $Z = Z_1, \dots U_i, \dots U_j, \dots Z_n$ having unknown words U_i and U_j. For a known word Z_i we simply hypothesize the source word itself. For an unknown word U_j, we propose a list of MSA candidates $\{A_1 \dots A_m\}$ using one or several methods discussed in the last section. Then we search for the best viterbi path $A = Z_1, \dots A_i, \dots A_j, \dots Z_n$ according to:

$$p(A|Z) = \text{argmax} \prod_{i}^{n} p(A_i|Z_{i-k+1} \ldots Z_{i-1}) \qquad (1)$$

where k is the order used for monolingual language model. We train a 5-gram language model. We use a stack-based search with a beam-search algorithm similar to that used in Pharoah [20] to select the best viterbi path. A large monolingual language model built on MSA data is used to score the candidates. The decoder decodes monotonically covering one Egyptian word at a time. For each Egyptian sentence we get a 1-best MSA sentence which is then decoded using MSA-to-English decoder.

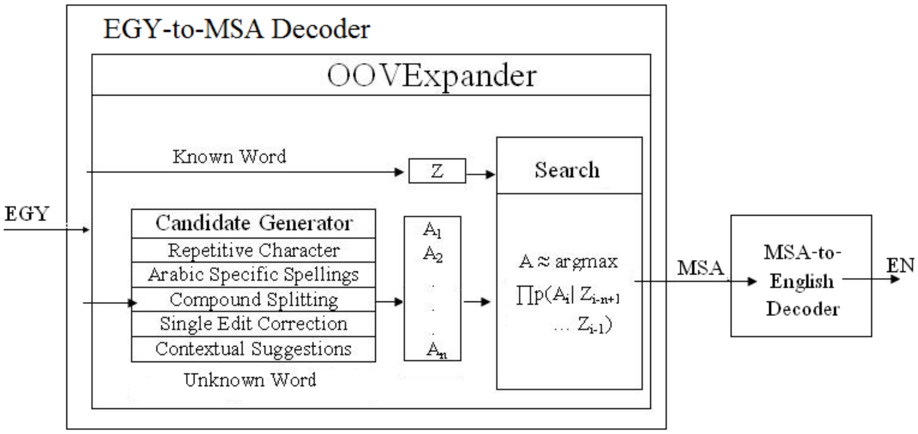


Fig. 2. Model for Egyptian-to-English Translation

4.2 Overall Model

The overall model for Egyptian-to-English translation is shown in Figure 2. Egyptian is first converted into MSA through an Egyptian-to-MSA decoder (as discussed in last section). The 1-best output is than passed to an MSA-to-English decoder. For the baseline system we bypass Egyptian-to-MSA decoding and translate Egyptian text using MSA-to-English decoder.

The decoding approach in [11] is superior to ours from the prospective that all MSA candidates of an Egyptian word are directly hypothesized into an MSA-to-English SMT system. In comparison our system proposes candidates and selects 1-best of these in a preprocessing step. The decision is based on just one monolingual language model feature. Because the language model is built on the large monolingual MSA data and the test set is Egyptian dialect, there is a domain mismatch and often time context does not prove to be useful to select the best candidates. In case of a tie, the decoder randomly picks one of the candidates. This is not the case in their system where n-best candidates are directly hypothesized in the MSA-to-English SMT, best English translation is chosen based on all features in the MSA-English MT model. Our approach, however, differs from the prospective that we propose MSA candidates for stems and not the affixes, whereas their work is only limited to affixes.

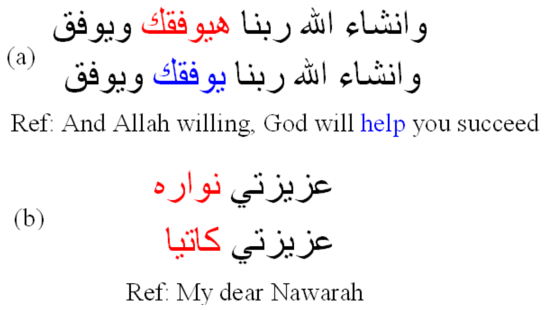
5 Results

In this section we discuss results obtained from running different proposed methods for OOV handling. For development purpose we test our edits to the baseline Egyptian input using TRL decoder because it is much efficient than the DTM decoder.

Table 4. OOV Handling Results – Applied To = Number of Egyptian words processing is applied to – Freq5 = Words having frequency $1 \leq 5$

System	TER	BLEU	Applied To
Baseline	60.40	19.49	0
A: Repeats	60.39	19.51	21
Dialect Rules	60.34	19.65	145
B: Spelling	60.14	19.95	510
C: Compounding	61.31	19.83	603
Synonyms	61.04	19.62	666
A + B	60.14	19.96	531
A + B + C	60.34	20.01	629
A + B + Freq5	59.96	20.01	863
A + B + C + Freq5	60.16	20.06	961

Our best component result (See Table 4) is obtained by the spelling correction module which hypothesize candidates that are at a single edit distance of the unknown word. Dialectal rules show small improvements because these are applied to a small number of words. Using compound splitting as a method of proposing candidates results in drop in TER score. Applying compounding to all OOV words hurts performance because most OOV words contain smaller components which have different meanings. Using synonym acquisition as a method of proposing MSA candidates also did not improve the results. Figure 3 shows examples of good and bad candidate proposed using this method. In Figure 3 (a) for the unknown word هيوفقك (help), and given the history وانشاء الله ربنا (God will), the language model propose all the candidate words that are actions, people expect from God such as يوفقك (help), يحميه (protect) , يرزقك (bless) etc. The language model score selects the right candidate in this case. An example of bad synonym is shown in Figure 3(b) where a list of proper names are proposed given the context عزيزتي (My dear), the language model selects another proper name كاتيا (Katia) rather than the correct candidate نوارة (Nawarah).



(a) وانشاء الله ربنا هيوفقك ويوفق
وانشاء الله ربنا يوفقك ويوفق
Ref: And Allah willing, God will help you succeed

(b) عزيزتي نواره
عزيزتي كاتيا
Ref: My dear Nawarah

Fig. 3. Examples of Good (a) and Bad (b) Synonym Suggestions

Table 4 also show different system combination that we attempted. In system $A+B+C$ and $A+B+C+Freq5$, compounding is applied only to those OOVs for which candidates are not proposed through other methods. In the last two rows we also propose candidates for the low frequency words appearing one to five times in the translation table. The results improve slightly.

Table 5. OOV Handling Results – Using DTM decoder

System	**Dev**		**Test**	
DTM	TER	BLEU	TER	BLEU
Baseline	56.88	23.87	59.86	22.00
A + B + Freq5	56.36	24.77	58.91	23.64
A + B + C + Freq5	56.51	24.85	58.95	23.72

Table 5 show gains obtained on dev and test set by running DTM decoder instead of TRL to test the preprocessed Egyptian input. Our front-end handling of OOV dialectal Egyptian words show an improvement of 1 BLEU point on dev and 1.7 BLEU points on the test set.

6 Mapping Egyptian to MSA through Pivoting

In this part of the paper we shift our focus towards mapping all Egyptian words, and not just the OOVs, into MSA. We use the well-known Convolution Model, previously used for an Arabic information retrieval task [21]. The idea is to pivot English as an informant between Egyptian and MSA. The model for mapping an Egyptian word Z to an MSA word A is given as:

$$p_c(A|Z) = \sum_{i}^{n} p(A|e_i)p(e_i|Z) \tag{2}$$

We use Model-1 to estimate the probability distributions $p(e|Z)$ and $p(A|e)$. The probability distribution $p(e|Z)$ is estimated from the 8.5K parallel sentences of Egyptian-English and $p(A|e)$ is estimated using 300K sentences of MSA-English. In order to overcome the sparse $p(e|Z)$ distribution we interpolate it with the Model-1 distribution $p_a(e|Z)$ built from the MSA-English corpus as:

$$p(e|Z) = \lambda p_z(e|Z) + (1 - \lambda)p_a(e|Z)$$

Equation 2 can then be rewritten as:

$$p_c(A|Z) = \sum_{i}^{n} p(A|e_i)[\lambda p_z(e_i|Z) + (1 - \lambda)p_a(e_i|Z)]$$

We select the top-10 MSA candidates according to the convolution model and search for the MSA sentences that maximizes the viterbi probability. The overall search (Equation 1, Section 4) is now redefined as:

$$p(A|Z) = \text{argmax} \prod_{i}^{n} p(A_i|Context)p_c(A_i|Z_i)$$

The results are shown in Table 6. Applying model to all Egyptian words in the dev set we notice a significant drop in the translation quality of Egyptian-to-English translation (See Z_1 in Table 6). Applying model to all Egyptian words in the dev set, hurts the translation quality because we are editing some words that our MSA-to-English system knows how to translate. In order to verify this hypothesis we tried to apply the model only to those Egyptian words that are frequent in the Egyptian-English data and less frequent in MSA-English data. $\mathrm{Z}_x\mathrm{A}_y$ in Table 6 means that candidates are proposed for an Egyptian word that appears at least x times in the Egyptian-English corpus and at most y times in the MSA-English corpus. However, results did not improve than the baseline system for any value of x and y. As the values for x and y are tightened i.e. model is only applied to only the words that are less frequent in the rich MSA-English corpus, we end up proposing candidates for only less than 450 words. If we try to apply model to only the Egyptian words that are seen at least 5 times in the Egyptian-English corpus (to avoid noisy alignments), the model is being applied to only 196 words and the results start converging towards the baseline.

The accuracy of the model in Table 6 is judged by the BLEU score of the MSA-to-English system. In order to scrutinize the results, we evaluated the accuracy of the

Table 6. Convolution Model – Proposed = Number of Egyptian words model is applied to and an MSA word different than Egyptian is selected in search

System	TER	BLEU	Proposed
Baseline	60.40	19.49	
Z_1	62.99	17.43	3361
Z_1A_{15}	60.82	19.18	450
Z_5A_{20}	60.59	19.33	196
$Z_{10}A_{500}$	60.66	19.29	362
$Z_{50}A_{100}$	60.53	19.41	62
UNK Word	60.48	19.50	80

convolution model in a more direct fashion. We used the 5000 most frequent Egyptian words list (also mentioned in Section 3.3). From this list we removed the Egyptian words that were translated to themselves. A remaining list of 1473 words is then used as a test set to evaluate the accuracy of the convolution model. The results are shown in Table 7. The 1-best and 10-best accuracies of the model are $\sim$7% and $\sim$22% respectively.

Table 7. Convolution Model Accuracy

	1-Best	10-Best
1473 Words	$\sim$7%	$\sim$22 %
5000 Words	$\sim$20 %	$\sim$50 %

In an another attempt to analyze the results we used the 100 sample sentences that were translated to MSA by human (Recall Section 1). In our results here we also measure the BLEU score at the intermidate step taking the human translated MSA as reference. The results of this controlled experiment is shown in Table 8. We see a slight improvement in both Egyptian-to-MSA and MSA-to-En systems, when applying the model to Z_3A_{20} i.e. Egyptian words that occured at least 3 times in the Egyptian-English corpus and at most 20 times in the MSA-English corpus. But the model is applied to only 50 words in this case. The BLEU score of 33.07 and TER score of 68.30 in the Egyptian-to-MSA baseline suggests that humans changed the Egyptian text significantly. However, when we try to apply the convolution model to all Egyptian words Z_1 we see a significant drop in the BLEU score.

We tried to analyze the output and found that most of the errors occur because of the sparse and noisy $p_z(e|Z)$ distribution built on the 8.5K Egyptian-to-English parallel data. The English informants proposed by the Egyptian word Z are incorrect. See Table 9 for examples.

7 Conclusion

In this paper we showed that the quality of Egyptian-to-English SMT can be improved by trying to map Egyptian to MSA, for which we have richer, more reliable translations. We proposed several methods to bridge the gap between Egyptian and MSA.

Table 8. Convolution Model – EGY-to-MSA = BLEU score for the edited Egyptian taking human translated MSA as reference, MSA-to-En = BLEU score for the edited Egyptian after decoding into English, Proposed = Number of Egyptian words model is applied to and an MSA word different than Egyptian is selected in search

System	**EGY-to-MSA**	**MSA-to-En**	**Proposed**
Baseline	33.07	15.22	
Z_1	24.88	12.74	808
Z_1A_{20}	32.23	15.14	123
Z_3A_{20}	33.27	15.34	50
Z_5A_{25}	33.23	15.29	42

Table 9. Convolution Model – Example Candidates Suggestion According to $p_z(e|Z)$

Word	**Correct**	**Suggested**	**Output**
دستوريه	Constitutional	Top	اعلي
ثوره	Revolution	Qadafi	القذافي
ربنا	Lord	Unbelievers	الكفار

We removed repetitions, applied Egyptian specific mappings, tried spelling correction, used compound splitting and suggested synonyms based on context etc. We also applied convolution model using English as a pivot to map Egyptian words into MSA. Our spelling-based correction showed improvement of 1.7 BLEU points over the baseline system, that translates unedited Egyptian into English.

Acknowledgments

We would like to thank the anonymous reviewers for their helpful feedback and suggestions. Nadir Durrani received funding from the European Union Seventh Framework Programme (FP7/2007-2013) under grant agreement n ° 287658. This publication only reflects the authors views.

References

1. Zbib, R., Malchiodi, E., Devlin, J., Stallard, D., Matsoukas, S., Schwartz, R., Makhoul, J., Zaidan, O.F., Callison-Burch, C.: Machine translation of arabic dialects. In: The 2012 Conference of the North American Chapter of the Association for Computational Linguistics, Montreal, Association for Computational Linguistics (2012)
2. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: Bleu: a method for automatic evaluation of machine translation. In: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics. ACL '02, Morristown, NJ, USA, Association for Computational Linguistics (2002) 311–318
3. Tillmann, C., Ney, H.: Word reordering and a dynamic programming beam search algorithm for statistical machine translation. Computational Linguistics **29** (2003) 97–133
4. Ittycheriah, A., Roukos, S.: Direct translation model 2. In: Proceedings of the North American Chapter of the Association for Computational Linguistics and Human Language Technologies Conference (NAACL-HLT), Rochester, NY (2007)
5. Snover, M., Dorr, B., Schwartz, R., Micciulla, L., Makhoul, J.: A study of translation edit rate with targeted human annotation. In: In Proceedings of Association for Machine Translation in the Americas. (2006) 223–231

6. Habash, N.: Introduction to Arabic Natural Language Processing. Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers (2010)
7. Chiang, D., Diab, M.T., Habash, N., Rambow, O., Shareef, S.: Parsing arabic dialects. In: EACL. (2006)
8. Habash, N., Rambow, O.: Magead: A morphological analyzer and generator for the arabic dialects. In: Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics, Sydney, Australia, Association for Computational Linguistics (2006) 681–688
9. Abo Bakr, H.M., Shaalan, K., Ziedan, I.: A hybrid approach for converting written egyptian colloquial dialect into diacritized arabic. In: Proceedings of the 6th International Conference on Informatics and Systems, INFOS2008, Cairo, Egypt (2008)
10. Sawaf, H.: Arabic dialect handling in hybrid machine translation. In: Proceedings of the Conference of the Association for Machine Translation in the Americas (AMTA), Denver, Colorado, Association for Machine Translation in the Americas (2010)
11. Salloum, W., Habash, N.: Dialectal to standard arabic paraphrasing to improve arabic-english statistical machine translation. In: Proceedings of the First Workshop on Algorithms and Resources for Modelling of Dialects and Language Varieties, Edinburgh, Scotland, Association for Computational Linguistics (2011) 10–21
12. Sajjad, H., Darwish, K., Belinkov, Y.: Translating dialectal arabic to english. In: Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), Sofia, Bulgaria, Association for Computational Linguistics (2013) 1–6
13. Durrani, N., Sajjad, H., Fraser, A., Schmid, H.: Hindi-to-urdu machine translation through transliteration, Uppsala, Sweden (2010)
14. Nakov, P., Tiedemann, J.: Combining word-level and character-level models for machine translation between closely-related languages. In: Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), Jeju Island, Korea, Association for Computational Linguistics (2012) 301–305
15. Al-Sabbagh, R., Girju, R.: Mining the web for the induction of a dialectical arabic lexicon. In Chair), N.C.C., Choukri, K., Maegaard, B., Mariani, J., Odijk, J., Piperidis, S., Rosner, M., Tapias, D., eds.: Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10), Valletta, Malta, European Language Resources Association (ELRA) (2010)
16. Riesa, J., Yarowsky, D.: Minimally supervised morphological segmentation with applications to machine translation. In Chair), N.C.C., Choukri, K., Maegaard, B., Mariani, J., Odijk, J., Piperidis, S., Rosner, M., Tapias, D., eds.: Proceedings of the 7th Conf. of the Association for Machine Translation in the Americas (AMTA 2006), Cambridge, MA, Association for Machine Translation in the Americas (AMTA) (2006)
17. Durrani, N., Hussain, S.: Urdu word segmentation. In: Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Los Angeles, California, Association for Computational Linguistics (2010) 528–536
18. Baroni, M., Bisi, S.: Using cooccurrence statistics and the web to discover synonyms in technical language. In: In Proceedings of LREC 2004. (2004) 1725–1728
19. Hagiwara, M., Ogawa, Y., Toyama, K.: Selection of effective contextual information for automatic synonym acquisition. In: Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics, Sydney, Australia, Association for Computational Linguistics (2006) 353–360
20. Koehn, P.: Pharaoh: A beam search decoder for phrase-based statistical machine translation models. In: AMTA. (2004) 115–124
21. Franz, M., McCarley, J.S.: Arabic information retrieval at ibm. In: TREC. (2002)