Building an English-Iraqi Arabic Machine Translation System for Spoken Utterances with Limited Resources

Jason Riesa[1], Behrang Mohit[2], Kevin Knight[3], Daniel Marcu[3]

[1]Department of Computer Science, Johns Hopkins University, Baltimore, MD 21218, USA
[2]Intelligent Systems Program, University of Pittsburgh, Pittsburgh, PA 15260, USA
[3]Information Sciences Institute, University of Southern California, Marina del Rey, CA 90292, USA

riesa@jhu.edu; behrang@cs.pitt.edu; {knight, marcu}@isi.edu

Abstract

This paper presents an English-Iraqi Arabic speech-to-speech statistical machine translation system using limited resources. In it, we explore the constraints involved, how we endeavored to mitigate such problems as a non-standard orthography and a highly inflected grammar, and discuss leveraging existing plentiful resources for Modern Standard Arabic to assist in this task. These combined techniques yield a reduction in unknown words at translation time by over 40% and a +3.65 increase in BLEU score over a previous state-of-the-art system using the same parallel training corpus of spoken utterances.

Index Terms: speech translation, limited resources, Arabic

1. Introduction

The Arabic spoken dialect of Iraq is a language deprived of the vast resources that researchers enjoy when working with its written counterpart, Modern Standard Arabic (MSA). While the intersection of vocabulary for Iraqi Arabic and MSA is substantial, at least 20% of the Iraqi Arabic lexicon is distinct from this set, having been heavily influenced by Persian, and, to a lesser extent, Turkish.

Despite large lexical and phonemic differences, Iraqi Arabic grammar remains distinctly Semitic. However, major grammatical deviations from MSA [1] include changes in word order, the absence of a noun declension system, and several modifications to the standard set of inflectional morphemes.

With these differences in mind, we describe the development of the required software components to build a mobile speech translation device in order to aid communication in urban situations between English-speaking military personnel and Iraqi Arabic-speaking civilians. Due to the nature of the domain and task, the input to the speech system can be expected to be noisy, with a high rate of profanities, disfluencies, and transcription errors in the training data, which should ultimately be corrected or eliminated.

2. Speech-to-Speech Translation for a Mobile Device

In a single direction, an Arabic-to-English speech-to-speech translation system requires an Arabic speech recognition component[1], an Arabic-to-English machine translation component[2], and finally

---

[1]Developed by SRI International, Inc.
[2]Developed jointly at USC/ISI and Language Weaver, Inc.
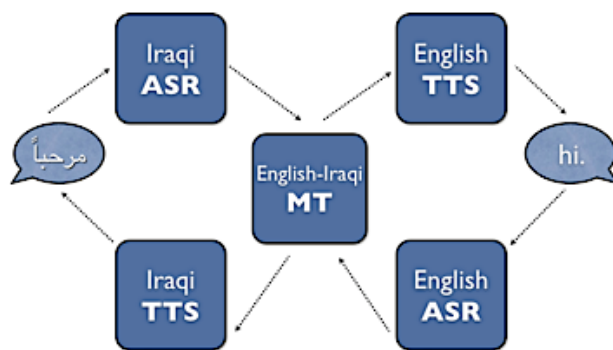

Figure 1: *System schematic for the bidirectional speech-to-speech translation system.*

an English text-to-speech component[3]. The components for the English-to-Arabic direction are defined analogously. Figure 1 shows the relationships among system components. Both Arabic and English speech recognition components generate textual output given their acoustic inputs, which is then passed to the machine translation component. The machine translation component translates the source text into the target language, and passes the translation on to the appropriate text-to-speech engine.

The entire system, composed of these components, is installed on a laptop, ruggedized for use in urban situations. In what follows, we concentrate on the machine translation component of this project. We first describe a baseline system, its evaluation using the BLEU score MT evaluation metric [2], and challenges limiting the performance of the system. Then, for each issue presented, we detail our solution and impact on BLEU. Finally, each technique described is combined into the final translation system to yield a substantial improvement in BLEU score and output quality.

We evaluate the techniques described below with respect to a baseline system: a state-of-the-art phrase-based statistical machine translation system described in [3]. We train this baseline system on our English-Iraqi Arabic parallel corpus of 36,895 utterances, and evaluate using a held-out test set of 1,903 utterances and one reference translation for each utterance. This yields a BLEU score of 26.07, with 95% confidence interval 24.97 - 27.42.

---

[3]Developed by Cepstral, LLC.

Corpus Statistic	English Training	Iraqi Training	English Test	Iraqi Test
Utterances	36,895	36,895	1,903	1,903
Running words	438,911	305,889	17,457	12,094
Words per utterance	11.9	8.3	9.2	6.4
Unique words	8,776	29,238	1,701	3,454

Table 1: *Summary statistics for parallel training and test corpora. The training corpus compiled from 40 hours of in-domain, transcribed and translated English and Iraqi speech audio as part of the DARPA Transtac program. A set of 17 dialogues was held out from the original training data to form a separate test set such that no speaker in the test set would appear in training. MT component evaluation is performed on transcriptions of these dialogues.*

3. A Highly Inflected Grammar

Iraqi Arabic is rich with prefix and postfix morphology. This likely accounts for the large difference between this and English in terms of utterance length and observed unique words. As seen in Table 1, the Iraqi vocabulary, without morphological analysis, is 3.33 times larger than its counterpart English vocabulary for the same training corpus. This means that much more parallel data is required in order to accurately learn translations for the increased number of unique words and phrases that result. Given the relatively small amount of parallel data for the task at hand, this presents an additional challenge for producing high-quality translations.

3.1. Morpheme segmentation

Given a list of known affixes and a list of known uninflected words, we employ a rule-based scheme to perform morpheme segmentation on Arabic words. In doing this, we are able to reduce the total vocabulary size by over 40%, and reduce the number of unknown words, words unable to be translated by the system, by almost half. Perhaps most importantly, this raises the BLEU score from 26.07 to 28.50, with 95% confidence interval 27.23 - 29.82. This is a 2.43 point increase over the baseline system.

Since the number of inflectional affixes is small, a list of known affixes can easily be compiled with the help of a basic grammar book, such as [4]. See Table 2 for a list of affixes used here and their respective glosses. We segment only those inflectional morphemes which often align quite naturally to their English translations.

To compile a list of known uninflected words, we make use of several resources. We first compile a list of all words seen in the training data not appearing to carry any known affix. More words are added to this list of stemmed words from a mined MSA dictionary as well as from manual examination of the data. Note that although the accuracy of the algorithm below is highly dependent on a reliable lexicon, it performs quite well with the wordlist compiled in a one person-day, as described above.

3.2. Segmentation algorithm

We assume an input Arabic word takes the form $p_1 p_2 ... p_n w s_1 s_2 ... s_m$, where w is inflected by n prefixes and m suffixes. For Arabic, $m = 1$. For each word in the corpus to be segmented, we first check to see if that word exists in the pre-compiled lexicon of uninflected words. If it does, then we skip segmentation and move on to the next word.

Prefix	Gloss		Suffix	Gloss
ال	the+		ي	+1-sg-pron
و	and+		ني	+1-sg-pron (verbal)
ل	for+		ك	+2-sg-pron
بـ	to/in+		ه	+3-sg-masc-pron
فـ	so/then+		ها	+3-sg-fem-pron
شـ	what+		نا	+1-pl-pron
ما	negation+		كم	+2-pl-masc-pron
مو	negation+		كن	+2-pl-fem-pron
لا	negation+		هم	+3-pl-masc-pron
للـ	for+the+		هن	+3-pl-fem-pron
هالـ	this+			
عالـ	on+the+			

Table 2: *Iraqi Arabic affixes considered by our morpheme segmentation method, not including over 20 combinations of these affixes found in the training corpus.*

Otherwise, if any combination of known affixes from Table 2 appear in the orthography, those affixes are segmented off to isolate w. The segmentation is kept if w appears in the training corpus. Otherwise, the input word is left as-is.

In the uncommon event that the segmentation is ambiguous, and multiple analyses are possible, we take the analysis with the most frequently occurring w, as counted in the training corpus. For our task and training corpus, this heuristic proved to yield reliable results in terms of choosing the correct segmentation. In other situations, more sophisticated methods may be required.

4. A Non-standard Orthography

A second challenge presented by the data given is widespread spelling errors and inconsistencies in both English and Iraqi Arabic. For example, Table 3 shows the counts and respective orthographies for the eight different ways in which the Arabic first-person singular pronoun was transcribed. On the English side, there exist many different transliterations for a single given Arabic proper name. For example, Qoran/Qor'an/Koran.

Traditionally, this problem is often mitigated by applying a global set of character-based normalization rules to a given text, e.g. "All instances of character X are mapped to character Y" for character pairs (X,Y) observed to be frequently interchanged. While this does have the effect of standardizing orthographies of many words, it may also introduce many potential ambiguities, as well.

As a concrete example, consider the following two pairs of words differing in edit distance by a single character:

	Gloss	Observed Arabic	Transliteration
(1)	name	إسم	<sm
(2)	name	أسم	>sm
(3)	imam	إمام	<mAm
(4)	in front of	أمام	>mAm

Suppose that every instance of the Arabic character إ, translit-

erated as (<), is normalized to the Arabic character آ, transliterated as (>). Then the two orthographies for the words meaning "name" become standardized, but a new ambiguity is introduced in words (3) and (4) after normalization.

For speech-to-speech applications, this method of normalization may be less than ideal. Characters which are eliminated or changed due to normalization, such as the Arabic *hamza* (ء), generally carry important acoustic information for a text-to-speech engine. Thus, the ultimate goal is to have a single canonical orthography for semantically identical words, and minimize the amount of new homographs introduced into the text.

The technique we describe attempts to standardize the orthography of words of type (1) and (2), while avoiding the introduction of new ambiguities which arise from context-independent, global character-based changes.

4.1. Algorithm

We would like to cluster groups of Arabic words, varying minimally in orthography, but all having identical semantics. Orthographic distance is computed using a weighted Levenshtein distance metric, measuring the minimum number of insertions, deletions, and substitutions of characters necessary to get from one orthography to another. In standard Levenshtein distance, each of these operations constitutes a cost of 1. We assign a *substitution* cost of 0 for any substitution among characters observed to be often interchanged. Not surprisingly, in our corpus these sets of characters are variations on the same base glyph, such as { ا أ إ آ }.

In addition, our scheme sets *deletion* costs of word-final characters ة and ه to 0, since the transcription of these was also not consistent throughout the text. Thus, the cost of the substitution أ → إ is 0, while the substitution cost of أ → ي remains 1.

After orthographic distance is computed, the existence of shared semantics among words in a given cluster, already determined to be similar in orthography, is decided via contextual analysis. For each word in the group, two vectors of word frequencies are computed. The first is a pre-context vector, holding counts of the k^- words preceding the word in question. The second is a post-context vector, holding counts of the k^+ words following the word in question. In our experiments, we set $k^- = k^+ = 1$.

Cosine similarity among all pre-context vectors is computed, and likewise for all post-context vectors. If all of the pre-context or post-context vectors (or both) are within a certain threshold of cosine similarity to each other, then the words in that group are deemed semantically equivalent.

For each group of words collected, whose members are judged to be semantically equivalent, we would like to normalize the orthography of each to a single standard form. We pick as the standard orthography the word w_{std} with highest frequency in the training corpus, and produce normalization rules $w_i \rightarrow w_{std}$ for each word $w_i \neq w_{std}$ in the group.

Thus, the object of this algorithm is to induce a list of normalization rules from the raw training corpus. After groups of words are collected, and standard orthographies for each group are chosen, we then apply those normalizations to the text in the training and test corpora.

	Orthography	Count
(a)	آني	5452
(b)	أنا	628
(c)	اني	464
(d)	آنا	414
(e)	انا	157
(f)	أني	38
(g)	آنی	30
(h)	انی	1

Table 3: *Observed orthographies and respective counts for the Iraqi Arabic first-person singular pronoun.*

5. Results for Segmentation and Orthographic Normalization

We report both cumulative and component BLEU scores. Cumulative BLEU scores were derived by cumulatively adding optimizations to a baseline system in decreasing order of impact. The degree of impact on the BLEU score was determined by the component BLEU scores, derived by evaluating the system with each individual vocabulary optimization in isolation.

Table 4 shows BLEU scores with confidence intervals for each vocabulary optimization technique described above, in addition to two operations performed on the English side of the parallel corpus: (1) Segmenting *'s* from all words, e.g. *Ali's → Ali 's*. This is generally standard practice. (2) Manually standardizing English transliterations of Iraqi proper nouns and applying spell-check to the entire corpus with *ispell*.

Combining these with morpheme segmentation and orthographic normalization as vocabulary optimization steps before training, yields a final BLEU score of 29.72. This is a +3.65 point increase over the 26.07 score for the baseline system.

Most of the improvement over the baseline can be attributed to Arabic morpheme segmentation. Indeed this technique resulted in the largest percent reduction in vocabulary size and number of unknown words. For the baseline system trained on the entire parallel corpus, there were 29,238 unique Iraqi Arabic words. After segmentation there were 17,138 – a 41.4% reduction. After orthographic normalization, vocabulary size was further reduced to 16,770 to give a 42.6% total reduction in Iraqi Arabic vocabulary size.

For the baseline system, there are 634 instances of unknown words after translation of the test set containing 12,094 Arabic words. After morpheme segmentation there are 356 unknown words. This is a total reduction in unknown words by 43.8%.

6. Revisiting Translation and Language Models

Our parallel corpus consists of two types of utterances: (1) utterances spoken by native Iraqi speakers (labeled *A*) and their English translations (labeled *A2E*), (2) utterances spoken by native English speakers (labeled *E*) and their Arabic translations (labeled *E2A*).

In the training corpus, Arabic utterances from native speakers and Arabic utterances derived from translated English differ in vocabulary and sentence structure. For example, many Ara-

Optimization Type	BLEU (Cumulative)	95% Confidence Interval	BLEU (Component)	95% Confidence Interval
Baseline	26.07	24.79 - 27.42	26.07	24.79 - 27.42
+Iraqi Segmentation	28.50	27.22 - 29.82	28.50	27.22 - 29.82
+English Segmentation	28.87	27.66 - 30.21	26.78	25.44 - 28.07
+Iraqi Normalization	29.33	28.09 - 30.63	26.69	25.38 - 28.06
+English Norm./Spelling	**29.72**	28.47 - 31.04	26.52	25.25 - 27.89

Table 4: *Cumulative and component BLEU scores. In order to derive cumulative BLEU scores, the optimization type with the greatest positive impact on the baseline system is added first. Then, subsequent optimizations are added in decreasing order of impact as denoted by component score.*

Customization	BLEU score
Baseline (no re-weighting)	27.84
LM only	28.93
TM only	28.40
LM and TM	29.17

Table 5: *Impact on BLEU score from applying three different re-weighting schemes to the Translation and Language models.*

bic idiomatic expressions only appear in the *A* labeled utterances. For translation of' these utterances, we can distribute extra weight to *A/A2E* utterance pairs for the translation and language models when the spoken input language is Arabic, and similarly for English and *E/E2A* utterance pairs. In effect, the translation and language models might be considered *customized* to the speaker.

We consider re-weighting the training data for both the language model (LM) and translation model (TM), and consider both in isolation. These are customizations for the scenario in which the spoken input language is Arabic; we expect similar results for English. In these experiments, the system is evaluated with a subset of the test set from Table 1, comprised only of the *A* utterances. Table 5 shows re-weighting of both LM and TM provide the largest positive impact.

7. Discussion

We have discussed several techniques aimed at increasing the accuracy of a machine translation system trained on limited resources. These techniques are easily implemented and incorporated into existing machine translation systems as preprocessing steps applied to a parallel corpus, and have been shown to provide benefit to overall output quality through significant BLEU score increases. Contextual orthographic normalization, in particular, requires little linguistic knowledge, save for setting a reasonable weighting scheme for the employed string-distance function. This technique can be easily adapted to other languages in which spelling or diacritical inconsistencies are common.

From the results of the preceding experiments, we make several key conclusions:

- Our lexicon-based morpheme segmentation provides a significant boost to BLEU score and machine translation output quality in a sparse data setting.

- Context dependent word-based orthographic normalization is a feasible and practical alternative to global character-based normalization, especially for applications in which it is preferable to preserve acoustic information encoded in a text.

- The assistance of existing tools for Arabic morpheme segmentation and analysis developed for Modern Standard Arabic [5], [6] is limited when applied to Iraqi Arabic, and likely when applied to other dialectal variants with significant divergence in grammar and lexicon, as well.

- The inclusion of a large Modern Standard Arabic news corpus provides no significant gains in BLEU score when evaluated on the test corpus. However, results from subsequent human evaluations of the system trained on both Iraqi data and a 500,000 sentence MSA corpus show that inclusion of MSA is relevant for translating utterances with words or phrases outside the intended domain, since users of these systems often stray onto the fringes of the domain.

8. Acknowledgments

9. References

[1] Altoma, S. J., "The Problem of Diglossia in Arabic: A Comparative Study of classical and Iraqi Arabic", *Harvard Middle Eastern Monograph Series*, Cambridge, MA., 1969.

[2] Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J., "BLEU: a Method for Automatic Evaluation of Machine Translation", In *Proceedings of the 40th Annual Meeting of the ACL*, pages 311-318, 2002.

[3] Koehn, P., Och F. J., and Marcu D., "Statistical phrase-based translation", In *Proceedings of HLT-NAACL.*, pages 127-133, 2003.

[4] Alkalesi, Y. M., "Modern Iraqi Arabic", *Georgetown University Press*, Washington, DC., 2001.

[5] Diab, M., Hacioglu, K., and Jurafsky, D., "Automatic tagging of Arabic text: From raw text to base phrase chunks", In *Proceedings of HLT-NAACL.*, 2004.

[6] Buckwalter, T., "Buckwalter Arabic Morphological Analyzer Version 2.0", *Linguistic Data Consortium*, catalog number LDC2004L02, 2004.