Challenges in Processing Colloquial Arabic

Alla Rozovskaya, Richard Sproat, Elabbas Benmamoun
Department of Linguistics, University of Illinois at Urbana-Champaign, Urbana, IL 61801, USA
{rozovska, rws, benmamou}@uiuc.edu

Processing of Colloquial Arabic is a relatively new area of research, and a number of interesting challenges pertaining to spoken Arabic dialects arise. On the one hand, a whole continuum of Arabic dialects exists, with linguistic differences on phonological, morphological, syntactic, and lexical levels. On the other hand, there are inter-dialectal similarities that need be explored. Furthermore, due to scarcity of dialect-specific linguistic resources and availability of a wide range of resources for Modern Standard Arabic (MSA), it is desirable to explore the possibility of exploiting MSA tools when working on dialects.

This paper describes challenges in processing of Colloquial Arabic in the context of language modeling for Automatic Speech Recognition. Using data from Egyptian Colloquial Arabic and MSA, we investigate the question of improving language modeling of Egyptian Arabic with MSA data and resources. As part of the project, we address the problem of linguistic variation between Egyptian Arabic and MSA. To account for differences between MSA and Colloquial Arabic, we experiment with the following techniques of data transformation: morphological simplification (stemming), lexical transductions, and syntactic transformations. While the best performing model remains the one built using only dialectal data, these techniques allow us to obtain an improvement over the baseline MSA model. More specifically, while the effect on perplexity of syntactic transformations is not very significant, stemming of the training and testing data improves the baseline perplexity of the MSA model trained on words by 51%, and lexical transductions yield an 82% perplexity reduction.

Although the focus of the present work is on language modeling, we believe the findings of the study will be useful for researchers involved in other areas of processing Arabic dialects, such as parsing and machine translation.

Keywords: language modeling, Arabic dialects

1. INTRODUCTION

Processing of Arabic dialects is difficult for several reasons. First, there are not many texts of spoken Arabic available. Second, dialect-specific electronic resources, such as annotated corpora, dictionaries, and parsers have not been developed. Finally, it is hard to develop resources for each dialect, since data transcription is expensive and time-consuming, and there is a whole continuum of Arabic dialects. By contrast, a lot of resources exist for MSA. We therefore wish to determine how one can use MSA data and resources in order to improve language modeling of Arabic dialects. We use the test set perplexity to evaluate the quality of a language model. Our study thus addresses the following question: is it possible to improve the quality of a language model for Colloquial Arabic through use of MSA data? The rest of the paper is structured as follows: first, we describe the corpora and the resources that we use. We then present

language modeling experiments with Egyptian and MSA data. We conclude with a brief discussion of the results.

2. RELATED WORK

The idea of using MSA data to improve language modeling of Arabic dialects has been investigated before. Kirchhoff et al. (2002) experiment with various techniques in an attempt to make use of MSA data to improve language modeling of Egyptian Arabic. In particular, they explore mixing Egyptian language model with MSA model. While they are able to find optimal weights that allow them to slightly reduce the perplexity of the held-out data, the technique has no visible effect on word error rate. Similarly, constrained interpolation, whose purpose is to limit the degree by which MSA model can affect the parameters of the Egyptian model, does not yield improvement. They also combine interpolation with text selection, namely, selecting those sentences in the MSA corpus that are closer in style to conversational speech. This approach attempts to overcome the genre difference of the Colloquial Arabic corpus and newswire data. Since none of the approaches was found to produce a positive effect, they conclude that Standard Arabic and Egyptian Arabic behave like two distinct languages.

The conclusion is supported by the following result (Kirchhoff, 2002): adding 300M words of MSA data to Egyptian training data increases the percentage of trigrams in the Egyptian test set that are also found in the language model from 24.5% to 25%, i.e. almost no increase in coverage is observed. Performing a similar experiment in English, we establish that there is more to the result than just the difference of genre and topic between newswire text and conversational data. We compute the percentage of trigrams found in the CallHome American English (Canavan, 1997a) evaluation set that are also found in the "in-domain" training data. This is 34%, higher than the 24.5% Kirchhoff et al. report for Arabic on comparably sized Egyptian CallHome data, though this is not surprising given the larger number of inflected word forms (compared to English) even in Colloquial Arabic. However, the addition of trigrams from 227 million words of North American Business (NAB) text raises this to 72.5%, a substantial reduction of "out-of-vocabulary" trigrams. What we observe is a substantially different behavior from what Kirchhoff et al. observed for Arabic. This might lead one to expect that linguistic transformations on MSA might have a greater chance of helping language modeling for Colloquial Arabic, than merely selecting MSA text that is more "in domain".

In Rambow et al. (2005), the idea of developing a part-of-speech tagger and a parser for Levantine dialect of Arabic through use of Standard Arabic data and resources is explored. In particular, an approach is described of adapting an MSA tagger enhanced with linguistic knowledge about the dialect. For parsing, three approaches are presented: sentence transduction, treebank transduction, and grammar transduction. All three approaches investigate the idea of adapting an MSA-style parser to Levantine Arabic. For example, in treebank and grammar transduction approaches, lexical substitutions and structural and syntactic transformations are applied to MSA Treebank sentences. These techniques yield a statistically significant reduction in error rate when compared to the performance of baseline naïve MSA parser on the dialectal data. These results are encouraging, as they indicate the possibility of using effectively MSA resources in order to develop resources for Arabic dialects.

3. DATA

3.1 Colloquial Data and Resources

3.11 CallHome corpus of Egyptian Colloquial Arabic

The CallHome corpus of Egyptian Colloquial Arabic (ECA) (Canavan, 1997b), is a collection of transcribed telephone conversations between native speakers of Egyptian Colloquial Arabic, and is divided into three parts: training, development, and testing. In our experiments, we use the training data (130K word tokens) and the development data (32K word tokens). The ECA corpus comes in two versions: romanized (with vowels) and Arabic script. Romanized orthography is phonemically based. Initially, the conversations were transcribed in romanized form, then converted to script via lookup-and replace procedure through the LDC Lexicon. In our experiments, the script version of the corpus is used.

3.1.2 Lexicon of Egyptian Colloquial Arabic

The ECA corpus is accompanied by the Lexicon of Egyptian Colloquial Arabic (LDC, 2002). The Lexicon contains 51202 entries, most of which come from the ECA corpus. Lexicon entries are keyed on their romanized form, and contain Arabic script representation of the word, its morphological analysis, the stem, as well as phonological, stress, and frequency information in the ECA corpus.

3.2 MSA Data and Resources

3.2.1 Arabic Gigaword

The Arabic Gigaword (Graff, 2003) corpus is a newswire corpus of Modern Standard Arabic. The corpus contains texts from four different sources:

- Agence France Presse (AFA) 97M tokens
- Al Hayat News Agency (ALH) 142M tokens
- Al Nahar News Agency (ANN) 143M tokens
- Xinhua News Agency (XIN) 18M tokens

3.2.2 Penn Arabic Treebank

The Arabic Treebank (ATB) (Maamouri, 2004) consists of three parts:

- Part 1: 140K words from Agence France Presse
- Part 2: 144K words from Al Hayat
- Part 3: 340K words from Al Nahar

ATB data files are morphologically analyzed using Buckwalter's analyzer (2002), which for a given word produces all possible morphological analyses. Analysis includes information about stem and affixes that comprise the word. Human annotators selected the correct part-of-speech analysis from the output of the analyzer. Additionally, ATB provides a mapping from the Arabic POS tagset to Penn English tagset, which allows to "collapse" several Arabic tags into one English tag, such as map all adjectives to one class. The treebank also contains syntactic representations of the newswire files.

4. EXPERIMENTS

Language modeling experiments are performed with the SRI Language Modeling Toolkit (Stolcke, 2002). All language models are trigram language models with Good-Turing discounting and Katz backoff for smoothing and with the "<unk>" word included in the training corpus. Data pre-processing includes removing non-alphabetic characters, diacritics and punctuation. The test set size in all experiments is 32K word

tokens. Unless otherwise specified, the AFA portion of the Arabic Gigaword corpus is used for MSA data.

4.1 Baseline language models

We refer to all language models trained on words as baselines. Our main baseline model is trained on Egyptian data. Table 1 gives the performance of the language model: the perplexity reduces slightly as more training data is used. This seems intuitively correct, since more data should allow for better parameter estimation.

We also compare our results to a word-based model trained using MSA data. This is because the perplexity of the test data given a model trained on Standard Arabic is significantly higher than that given a model trained on Egyptian data, and while it might not be possible to reduce the latter by applying simple techniques, we would still like to evaluate the effect of each of those techniques by comparing against word-based language models trained on MSA data. Since we only have 130K words of training data for Egyptian available, we build an MSA model of the same size. Table 2 compares the performance of the MSA and ECA models on Egyptian data. The perplexity of the MSA model is about 65 times higher than that of the ECA model.

Training size	**Perplexity**
100K	188.7
130K	184.8

TABLE 1: Perplexity of ECA (word) on ECA (word)[1]

In order to get a sense of the difficulty of the task, we train a model on the MSA data and evaluate it using data from the same domain. It turns out that on in-domain data a perplexity of 955.4 is obtained, in contrast to 184.8 for ECA. We conjecture that one of the reasons for the high perplexity is the morphological complexity of Standard Arabic. We compute the vocabulary sizes of Egyptian and MSA data sets of 130K tokens. As shown in Table 3, MSA corpus has more than twice as many word types and 1.5 times more bigram types than the ECA corpus of the same size.

Training corpus	**Perplexity**
ECA	184.8
MSA	12874.2

TABLE 2: Performance of Egyptian and MSA corpora on Egyptian data[1]

	ECA	MSA
Vocabulary size	13,500	30,000
Number of bigrams	60,000	95,000

TABLE 3: Vocabulary and bigram comparison of ECA and MSA corpora

In order to determine whether increasing MSA training set size results in a better model, we train models with more data and evaluate them on in-domain and out-of-domain test sets. Perplexities are measured by varying the training set size from 130K to 27M word tokens. Figures 1 and 2 display the perplexity of ECA and MSA test sets, respectively, as a function of training set size. No perplexity reduction of the ECA data is observed. In fact, the perplexities increase with the increase of training set size. We believe the increase in perplexity is due to the fact that a larger training set has a bigger vocabulary and consequently has more parameters, so that less probability mass is

assigned to a single unseen event. Therefore the backoff probabilities are smaller in a larger model. The increase in perplexity thus indicates that adding more MSA data does not contribute to larger coverage of the ECA test set.

To verify the hypothesis that increased perplexities are simply caused by smaller backoff probabilities, we train a model on the English New York Times (NYT) text and evaluate it with the ECA test set. Surely, one would not expect the coverage of the English model on Egyptian data to improve as more training data are added. Figure 3 shows the behavior of MSA models from four different domains on the ECA test set and the behavior of the New York Times (NYT) model on the same test set. While ANN and ALH models outperform the others, all the corpora behave similarly in that their prediction ability does not improve as more data are added. By contrast, when tested on in-domain data, the performance of the MSA model improves consistently with more training data. Figure 2 illustrates that, supporting the idea that a correct language model should exhibit an analogous behavior. The perplexities reduce from 955.4 to 157.4 as the training set is increased from 130K to 27M word tokens.

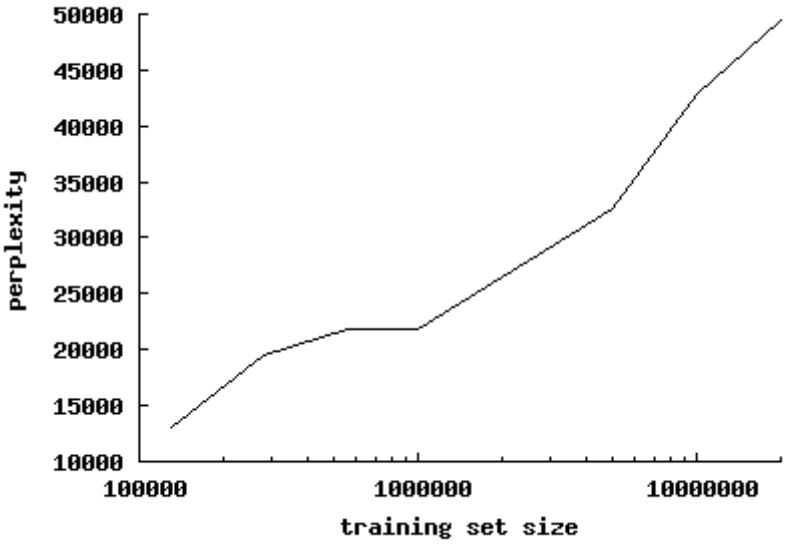

FIGURE 1: Performance of MSA model on ECA data as a function of training set size

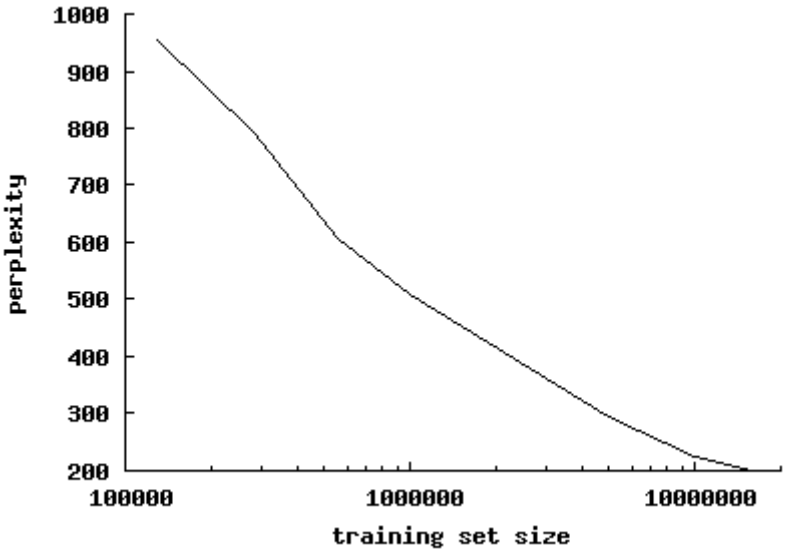

FIGURE 2: Performance of MSA model on in-domain data as a function of training set size

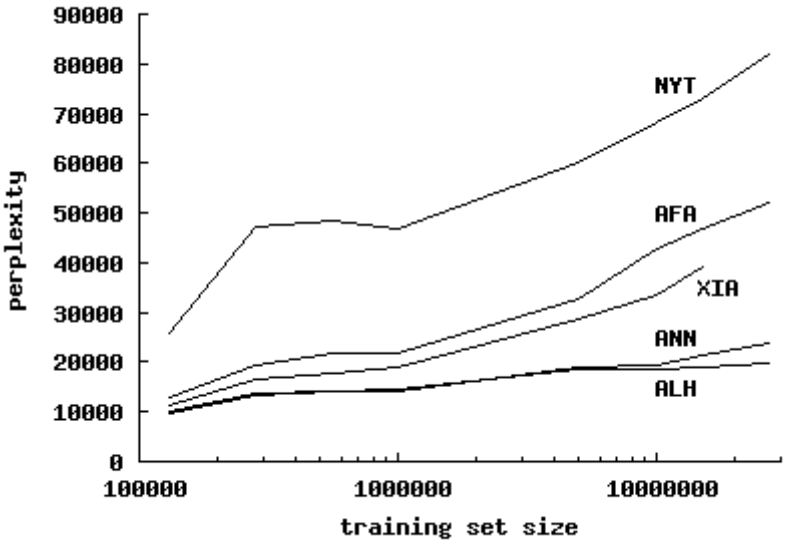

FIGURE 3: Performance of New York Times and four MSA models on ECA

4.2 Stem language models

The main assumption behind a stem language model is that removing inflections will reduce the amount of morphological discrepancy between the two dialects and will allow us to better model the spoken language with Standard Arabic data. The procedure consists of separating clitics (prepositions, determiners, direct object pronouns) and stripping affixes. We stem both the training and the test sets. Here is an example of a word where stemming is applied:

وسيجدنا ⟵ و جد نا

Gloss: and + will + he/it + find + us

The stem models for the MSA data are constructed using Buckwalter's Morphological analyzer and ASVM package that performs tokenization and POS tagging on Modern Standard Arabic (Diab, 2004). Specifically, for each word in the corpus, a list of candidate analyses is obtained using Buckwalter's morphological analyzer. We use the ASVM package to separate clitics and to select the correct POS tag for each token. The POS information is then used to select the correct stem out of the analyses produced by the morphological analyzer[2]. Stems for Egyptian data are obtained from the LDC Lexicon. In the case of multiple morphological analyses, one analysis is selected randomly.

We compute the vocabulary and bigram overlap of MSA and ECA data sets of 130K words. The proportion of word types and bigram types in the ECA corpus that are also found in the MSA corpus increases through stemming from 14% to 25.5%, and from 0.8% to 3.6%, respectively. It should be noted that while the proportion of bigrams increases more than four times due to stemming, the overlap is still very small. Table 4 gives the performance of the ECA and MSA models on in-domain data. Stemming reduces the perplexity by about 50% and 85%, respectively. Similarly, stemming leads to a 50% perplexity reduction when the MSA model is tested on ECA data (Table 5). Figure 4 displays the performance of the MSA model with respect to the ECA test set, as a function of training set size: as the training size increases, the perplexity also increases.

Training data	Testing data	Perplexity
ECA (word)	ECA (word)	184.8
ECA (stem)	ECA (stem)	89.1
MSA (word)	MSA (word)	955.4
MSA (stem)	MSA (stem)	140.4

TABLE 4: Perplexity of MSA and ECA models on in-domain data before and after stemming

Training data	Testing data	Perplexity
MSA (word)	ECA (word)	12874.2
MSA (stem)	ECA (stem)	6260.7

TABLE 5: Perplexity of MSA models of comparable sizes on ECA test[1]

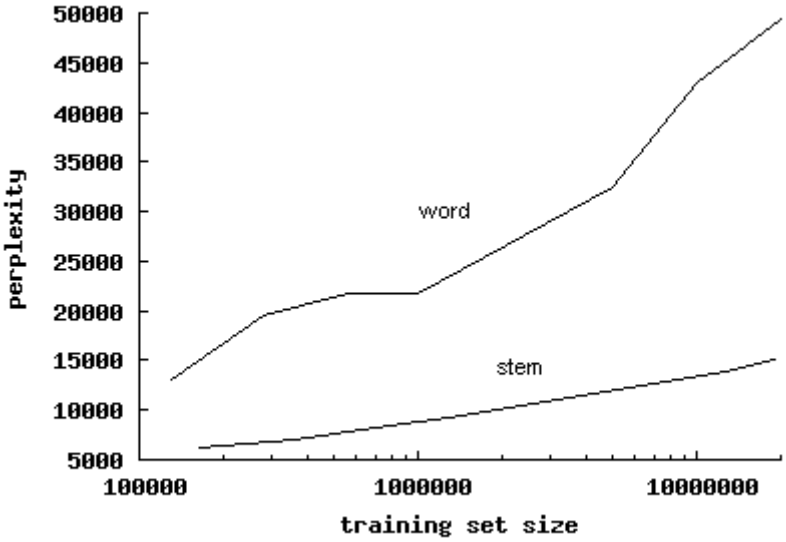

FIGURE 4: Performance of MSA model on ECA data as a function of training set size

4.3 Lexical Transductions

The idea of using dialect-to-Standard translations to improve part-of-speech tagging and parsing of dialectal Arabic is explored with a slightly different flavor in Rambow et al. (2005). We compile a list of all words occurring in the ECA corpus with frequency two or more and manually create a lexicon that specifies for each Egyptian word its MSA equivalent(s). The purpose of the lexicon is to account for words that are derived from different roots in the two dialects, share the same root, but display different morphological processes, or just have different spellings. Table 6 gives several sample lexicon entries.

We replace every word in the ECA data set (30K word tokens) with the corresponding MSA equivalent(s) specified in the lexicon. Each Egyptian word found in the lexicon is mapped to a list of stems of its MSA equivalents. Stems for MSA words are obtained using Buckwalter's morphological analyzer. About 82% of all word tokens in the ECA data set are found in the lexicon. For the rest of the words, stems specified in the LDC lexicon are used. We build a transducer for each ECA sentence, where every word is represented as a lattice of possible MSA stems. The stem MSA language model as described in Section 4.2 is used to obtain the most likely sequence of stems in the ECA sentence. We use fsmtools (Mohri) to find the most likely path in the lattice. This method reduces perplexity further from 6260.7 to 2262.3. The results are shown in Table 7.

Frequency in ECA corpus	ECA word	MSA equivalent(s)
967	هى	هي
628	احنا	نحن
8	جبتها	أحضرتها

TABLE 6: Sample Lexicon Entries

Baseline	12874.2
Stem	6260.7
Stem + Lexical Transductions	2262.3

TABLE 7: Performance comparison for different MSA models on ECA data

4.4 Syntactic transformations

This approach attempts to account for systematic syntactic differences between MSA and Egyptian dialects. The idea is similar to syntactic transformations described in Rambow et al. (2005), but we wish to discover such transformations automatically.

Using the Al Hayat part of the Arabic Treebank, we identify frequent syntactic productions in the corpus. We replace all terminal nodes in a tree with corresponding part-of-speech tags and map those tags to the English tagset in order to reduce the overall number of distinct subtrees. We compute frequency of every subtree type in the corpus, and select fifty with highest frequency. Every possible permutation of the child nodes for each frequent production $A \rightarrow B_1 B_2 \ldots B_n$, where B_i is any terminal or non-terminal symbol, is considered a transformation. We then apply every transformation to the Al Hayat Treebank corpus in order to determine which transformations are useful. A transformation is considered useful if its application to the training corpus leads to a perplexity reduction on the ECA test set. In this manner, we find about fifty useful transformations, each resulting in a 5-8% reduction of perplexity. However, when compared with the methods described above, the effect on perplexity of the transformations is not significant. Figures 5 and 6 show two examples of useful transformations. The child nodes participating in the transformation are in bold.

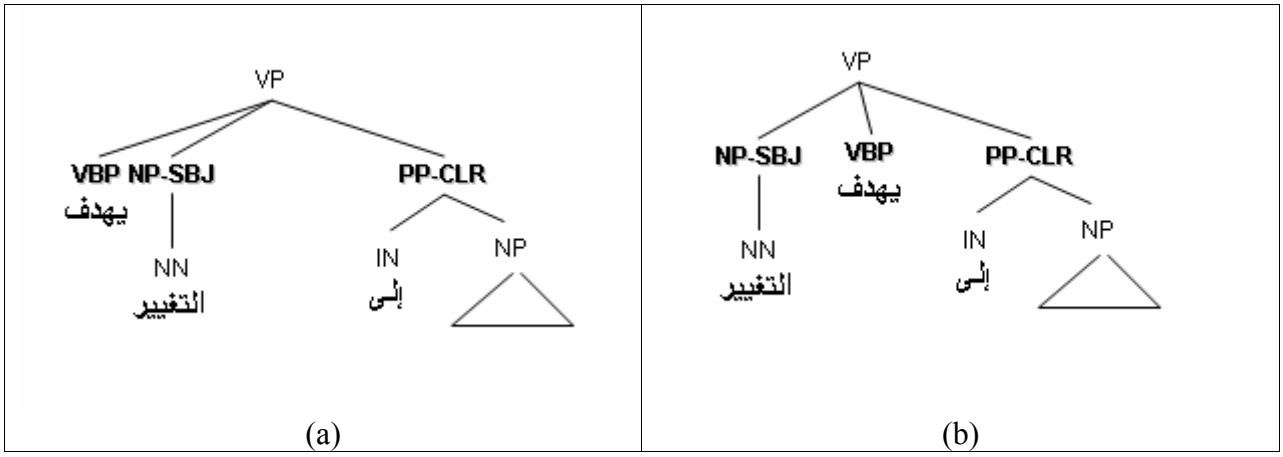

FIGURE 5: Example of syntactic transformation

FIGURE 6: Example of syntactic transformation

5. DISCUSSION AND CONCLUSIONS

We have described a variety of techniques that allowed us to improve the perplexity of the baseline (word) language model trained on MSA when tested on Egyptian Colloquial Arabic. However, we have not been able to improve over the model trained on Colloquial data. Furthermore, the general tendency of perplexity increase with increase of training set size remains. As mentioned in Section 4.1, we believe that a correct language model should display reduction in perplexity as training set increases, since parameters learned are more reliable. By contrast, an increase in perplexity may indicate that the training data are very different from the test data.

In light of the present experiments as well as the results obtained previously (Kirchhoff, 2002), we conclude that using MSA data does not help improve language modeling for Colloquial Arabic. However, since we have only experimented with Egyptian Arabic, more research is needed to determine whether the results that the present study has demonstrated hold across other dialects of Arabic.

ACKNOWLEDGEMENTS

We would like to thank Hala Jawlakh for preparing the Egyptian lexicon and for helping with numerous language-related questions. This research is supported by an NSF grant IIS 04-14117.

ENDNOTES

[1] Perplexities are computed by averaging the results of five runs. For each run, training and testing sentences are selected randomly from the corpus

[2] Since the Buckwalter Analyzer provides a much finer morphological analysis, complete disambiguation cannot be achieved with ASVM package, but allowed us to disambiguate with respect to stem about 96% of all tokens in the AFA corpus.

REFERENCES

Buckwalter, Tim (2002). *Arabic Morphological Analyzer Version 1.0.* Linguistic Data Consortium,Philadelphia

Canavan, Alexandra, George Zipperlen and David Graff (1997a). *CALLHOME American English Speech.* Linguistic Data Consortium, Philadelphia

Canavan, Alexandra, George Zipperlen, and David Graff (1997b). *CALLHOME Egyptian Arabic Speech.* Linguistic Data Consortium, Philadelphia

Diab, Mona, Kadri Hacioglu and Daniel Jurafsky. (2004). 'Automatic Tagging of Arabic Text: From raw text to Base Phrase Chunks'. In *Proceedings of HLT-NAACL 2004.*

Graff, David (2003). *Arabic Gigaword.* Linguistic Data Consortium, Philadelphia

Kirchhoff, Katrin et al. (2002). *Novel Speech Recognition Models for Arabic.* Johns-Hopkins University Summer Research Workshop 2002 Final Report

LDC, et al. (2002). Egyptian Colloquial Arabic Lexicon. Linguistic Data Consortium, Philadelphia

Maamouri, Mohamed et al. (2004). *Arabic Treebank: Part 2 v 2.0.* Linguistic Data Consortium, Philadelphia

Mohri, Mehryar , Fernando C. N. Pereira and Michael D.Riley. *Finite-State Machine Library.* http://public.research.att.com/~fsmtools/fsm/

Rambow, Owen, David Chiang, Mona Diab, Nizar Habash,Rebecca Hwa, Khalil Sima'an, Vincent Lacey, Roger Levy, Carol Nichols, and Safiullah Shareef (2005). *Parsing arabic dialects.* Final Report, 2005 JHU Summer Workshop.

Stolcke, A. (2002). 'SRILM- an extensible language modeling toolkit'. In *Proc. Int. Conf. on Spoken Language Processing*, Denver, Colorado, September 2002.