# Diacritization as a Machine Translation Problem and as a Sequence Labeling Problem

Tim Schlippe

ThuyLinh Nguyen

Stephan Vogel

23 October 2008

Universität Karlsruhe (TH)

Carnegie Mellon

## Outline

1. Motivation
2. Data Format
3. The Evaluation System
4. Diacritization as a Machine Translation Problem
   - The Baseline Systems (word level, character level)
   - Lexical Scores (beside relative phrase frequencies in phrase table)
   - The System on both Levels
   - The Post-Editing System
5. Diacritization as a Sequence Labeling Problem
   - Part-of-Speech Tagging
   - Conditional Random Fields
6. Conclusion and Future Work

# Motivation

interACT

## Ambiguity in Arabic

- Modern Arabic text normally composed of scripts without diacritic marks

| without diacritics | with diacritics | meaning | pronunciation |
|---|---|---|---|
| علم | عِلْم | science, learning | Eilm |
| | عَلَم | flag | Ealam |

## A diacritization system may …

- simplify text-to-speech and speech-to-text applications [Zitouni et al. 2006] [Zakhary 2006]
- improve translation Arabic → other language (e.g. passivation diacritic „damma“) [Diab et al. 2007]
- improve translation other language → Arabic (e.g. double case endings) [Gharieb 2006]
- benefit non-native speakers and sufferers of Dyslexia [Elbeheri 2004]
- be applied to other languages that also have diacritics that could lead to ambiguity – due to statistical features (e.g. Hebrew, Romanian, French) [Tufiş et al 1999] [Gal 2002]

## Buckwalter Transliteration

- To process data morphologically
- From Unicode and back it is a one-to-one mapping without any gain or loss of ambiguity

| Name | | Buckwalter Transliteration | Pronunciation |
|---|---|---|---|
| **Short vowels** /a/, /u/, /i/ | | | |
| Fatha | ◌َ | a | /a/ |
| damma | ◌ُ | u | /u/ |
| kasra | ◌ِ | i | /i/ |
| **Double case ending** | | | |
| fathatayn | ◌ً | F | /an/ |
| dammatayn | ◌ٌ | N | /un/ |
| kasratayn | ◌ٍ | K | /in/ |
| **Syllabification marks** | | | |
| shadda | ◌ّ | B (normally ~) | consonant doubling vowel |
| sukuun | ◌ْ | o | vowel absence |

# The Evaluation System

interACT

## Sclite

- Part of NIST Speech Recognition Toolkit
- Finds alignments between reference and hypothesis word strings

- Word Error Rate (WER)
  - with final vowelization (final_vow)
  - without final vowelization (no_final_vow)
- Diacritization Error Rate (DER)
  - with final vowelization (final_vow)
  - without final vowelization (no_final_vow)

➤ Distinction in final vowelization: analyze errors in stems and endings

➤ Distinction in WER and DER: operating on word and char level

# Diacritization as a Translation Problem

## Translation Process

- Monotone translation from undiacrized text to diacritized text
- Translate phrases by CMU SMT system [Vogel et al., 2003]

- **Translation on word level:**

| | | | |
|---|---|---|---|
| without diacritics | mwskw | Jf | b |
| with diacritics | muwsokuw | Jaf | b |

- **Translation on character level:**

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| m | w | s | k | w | space | J | f | space |
| mu | w | so | ku | w | space | Ja | f | space |

  – Split undiacritized text into individual consonants
  – Split diacritized text into consonant-vowel compounds
  – Insert special word separator to be able to restore words

# The Baseline Systems

## Data: LDC‘sTreebank of diacritized An Nahar News stories

- Training data: each 613 k words, 23 k sentences
- Dev data / Test data: each 32 k words, 2 k sentences

- No punctuation marks included
- Diacritics deleted to create undiacritized part of parallel corpus
- Used for
  - machine translation experiments except post-editing
  - sequence labeling experiments

# The Baseline Systems

## The Word Level System

- 10-gram Suffix Array Language Model
- Phrase table contains up to 5-gram entries and appropriate relative phrase frequencies
- Drawback: unknown word leads to word error

## The Character Level System

(according to [Mihalcea 2002])

- 10-gram Suffix Array Language Model
- Phrase table contains up to 5-gram entries and appropriate relative phrase frequencies
- All words can be diacritized:
  Each consonant is assigned to the same consonant with a diacritic
- Drawback: much less context is covered, e.g.

3-gram on character level:

| | | |
|---|---|---|
| m | w | s |
| mu | w | so |

3-gram on word level:

| | | |
|---|---|---|
| mwskw | Jf | b |
| muwsokuw | Jaf | b |

# The Baseline Systems

## Results of the Baseline System

| | | word-based | char-based |
|---|---|---|---|
| final_ vow | WER | 22.8 | 21.8 |
| | DER | 7.4 | 4.8 |
| no_final_ vow | WER | 9.9 | 7.4 |
| | DER | 4.3 | 1.8 |

- Better results with character level system
  since the word level system was not able to translate many words

  → First focus on the character level system

# Lexical Scores

## Additional Lexical Scores beside Phrase Translation Probabilities

- Relative frequencies unreliable for low frequency events ➤ Lexical scores
- Moses Package [Koehn et al., 2007] and GIZA++ [Och and Ney, 2003] to create phrase table with lexical scores beside relative frequencies, by default containing up to 7-gram entries
- Given a source phrase $f_1 \ldots f_J$ and a target phrase $e_1 \ldots e_I$ , we calculate:

$$lex(f_1^J|e_1^I, a) = \prod_{j=1}^{J} \underbrace{\frac{1}{|\{i|(j,i) \in a\}|}}_{= 1^*} \sum_{(j,i)\in a} w(f_j|e_i)$$

* alignment strictly monotone and one-to-one

➤ WER improvement by up to 7-gram phrases compared to char level baseline system: 0.2%

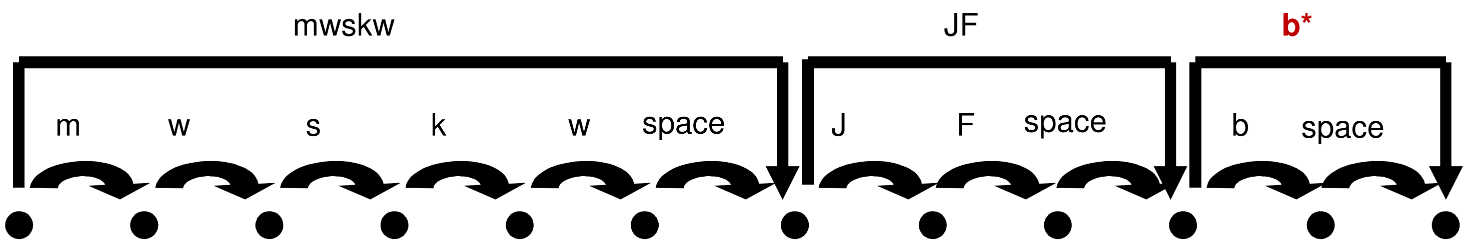
| | | baseline system | max. phrase length 7 | lexical score |
|---|---|---|---|---|
| final_vow | WER | 21.8 | 21.6 | 21.5 |
| | DER | 4.8 | 4.8 | 4.7 |
| no_final_vow | WER | 7.4 | 7.5 | 7.4 |
| | DER | 1.8 | 1.9 | 1.8 |

➤ Further WER improvement by lexical scores: 0.1%

# The System on both Levels

## Edges from Character to Character and from Word to Word

- If word known, use word level; otherwise go to character level




Lattice input with edges from character to character and from word to word (one char words marked)

| | | | |
|---|---|---|---|
| Word-part | mwskw | # mu w su ka w space |
| | J f | # Ja f space |
| | b* | # b space |
| | m w | # mu w |
| | … | |
| Char-part | b | # b |
| | J f | # Ja f |
| | J f space | # Ja f space |

Extract from the phrase table of the hybrid approach with word part and character part

- Due to the phrase count feature in the decoder translations from fewer phrases are preferred → bias towards edges from word to word
- LM still on character level ➤ next step: integrate word level LM

# The System on both Levels

## Integrating Word Level Language Model

- Generate 1000-best list for each sentence
- Convert from char representation to word representation
- Calculate language model score for each sentence
- Rescoring and reordering

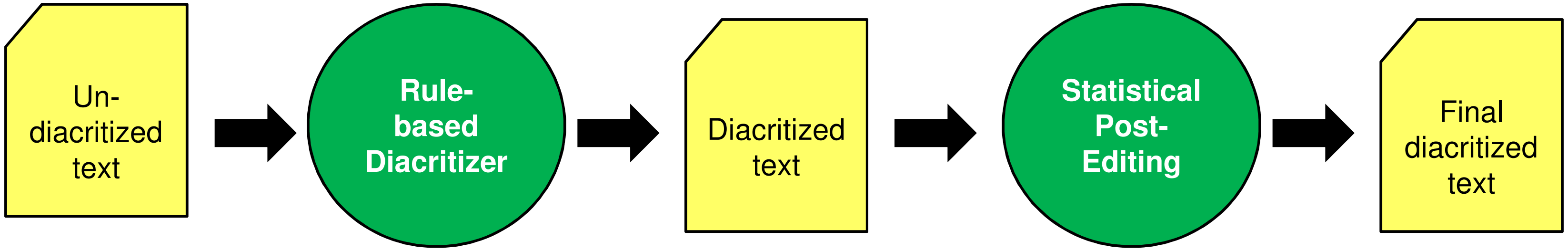
- Experiments with longer n-grams in the Suffix Array Language Model Toolkit [Zhang, 2006] as well as with the SRI Language Model Toolkit [Stolcke, 2002]

| language model | | char 5 | word 3 SRI | word 4 SRI | word 6 SA |
|---|---|---|---|---|---|
| final_vow | WER | 20.1 | 19.9 | 20.0 | 20.0 |
| | DER | 4.3 | 4.3 | 4.3 | 4.3 |
| no_final_vow | WER | 6.6 | 6.8 | 6.9 | 6.9 |
| | DER | 1.6 | 1.7 | 1.7 | 1.7 |

- WER improvement compared to system on character level: 0.9%
- WER improvement by word level LM: 0.2%
- No further improvement with longer n-grams

# The Post-Editing System

## Post-Editing the Output of AppTek’s Rule-Based Diacritizer




- Rule-based system excludes a large number of possible forms [Simard et al. 2007]
- For Post-Editing: Phrase table with phrase translation probabilities and lexical scores in both directions, created by Moses/GIZA++

# The Post-Editing System

interACT

## Data: Output of Rule-based System, Human Reference

- Training data: each 104 k words, 36 k sentences
- Dev data / Test data: each 6 k words, 2 k sentences

- As sentences are more similar and rather short, error rates with AppTek's data are lower than those obtained with LDC's Arabic Treebank data

## Results of the Post-Editing System

| | | baseline | post-editing |
|---|---|---|---|
| final_ vow | WER | 15.6 | 13.8 |
| | DER | 5.5 | 4.9 |
| no_final_ vow | WER | 10.3 | 9.3 |
| | DER | 3.5 | 3.2 |

WER improvement by: 0.8%

# Diacritization as a Sequence Labeling Problem

interACT

## Idea

- Errors at the word ending significantly higher than at the word stems
- Goal: integrate more global features and grammatical information
  - Conditional random fields

## Sequence Labeling

- Undiacritized word represented as a sequence of characters $X$
- We label each consonant in $X$ with none, one or more diacritics which should follow that consonant in diacritized form
- Task of diacritization of $X$ : Finding its sequence $Y$

| | | | | | | |
|---|---|---|---|---|---|---|
| mwskw | X: | m | w | s | k | w |
| | | \| | \| | \| | \| | \| |
| muwsokuw | Y: | u | $\epsilon$ | o | u | $\epsilon$ |

# Conditional Random Fields

interACT

## Conditional Random Fields

- Conditional random fields (CRFs) successful in parts-of-speech tagging and noun phrase chunking [Lafferty et al., 2001]
- The CRF model estimates the parameters $\overline{\theta}^*$ to maximize the conditional probability of the sequence of tags given the sequence of the consonants in the training data $\mathbf{T}$ as given by the following equation:

$$\overline{\theta}^* = \underset{\overline{\theta}}{\operatorname{argmax}} \sum_{(X,Y)\in\mathbf{T}} \log p\left(Y|X,\overline{\theta}\right)$$

where $\log p\left(X|Y,\overline{\theta}\right) = \sum_i \theta_i f_i\left(X_q, Y_q\right)$

$f_i$ feature function

$X_q, Y_q$ sub-sequences of $X, Y$

- At the test time, given a sequence of consonants $X$ and parameters $\theta^*$ found at the training time, we decode $X$ into the sequence $Y^*$.

$$Y^* = \underset{Y}{\operatorname{argmax}}\, p\left(X|Y,\overline{\theta}^*\right)$$

# Conditional Random Fields

## Parts-of-Speech

- apply CRF++ to assign the diacritics to the consonants on char level [Kudo, 2007]
- integrate grammatical information
  (identification of words as adjective, imperfect verb, passive verb, …; relationship with other words)
- Tags by Stanford Arabic Tagger (Penn POS Tags) [Toutanova and Manning, 2000]

| | | |
|---|---|---|
| waJawoDaHa | VBD | perfect verb |
| AlbaronAmaji | DTNN | determiner/demonstrative pronoun, common noun |
| AlBaCiy | WP | relative pronoun |
| yunaZBimu | VBP | imperfekt verb |
| muLotamarAF | NN | common noun |
| duwaliyBAF | JJ | adjective |
| yabodaJu | CD | cardinal number |
| JaEomAlahu | CD | cardinal number |

Example for POS Tags in Arabic

# Conditional Random Fields

## Results for different amounts of data and different context

- Output sequence dependent
  - on previous, current and following characters,
  - on the previous, current and following word
  - on parts of speech of previous, current and following word
- Problem: CRF++ requires a lot of memory
- Due to memory limitations trade-off between training corpus size and number of features

| | data | 100% | 75% | | | | |
|---|---|---|---|---|---|---|---|
| | context | 4 | 4 | 6 | 8 | 10 | 12 |
| final_vow | WER | 22.8 | 24.1 | 22.6 | 22.2 | 22.0 | 21.9 |
| | DER | 5.1 | 5.4 | 4.9 | 4.8 | 4.7 | 4.7 |
| no_final_vow | WER | 9.4 | 10.0 | 8.5 | 8.3 | 8.3 | 8.4 |
| | DER | 2.2 | 2.4 | 2.0 | 1.9 | 1.9 | 1.9 |

# Conclusion and Future Work

## Conclusion

- **Techniques from phrase-based translation**

  Improvements by:
  - Using longer phrases in the phrase table
  - Adding lexical scores in the phrase table
  - Operating both on word and character level
  - Rescoring with word-level LM

- **Sequence labeling by using conditional random fields** to integrate additional features like parts of speech
  - Due to memory limitations trade-off between training corpus size and number of features
  - We expect that with more data and additional features this approach will perform on the same level or better than translation approach

- **Post-Editing rule-based diacritizer with statistical system** outperformed both rule-based and pure statistical system

## Conclusion

- Major problem in diacritization are the errors in the word endings, e.g. in phrase-based diacritization systems word ending „pi“ (ta marbouta with kasra) occurs almost 2% and “i” (kasra) even more than 5.5% more frequently in our hypothesis than in the reference or in the training data

Distribution of the Word Endings in the

| Hypothesis of the Hybrid System with word LM | | Human Reference Translation | | Training Data | |
|---|---|---|---|---|---|
| pi | 10.477 | pi | 8.508 | pi | 8.828 |
| y | 6.876 | y | 6.890 | y | 7.122 |
| A | 6.477 | A | 6.432 | A | 6.252 |
| n | 4.906 | n | 4.956 | n | 4.716 |
| Y | 4.459 | Y | 4.436 | Y | 4.398 |
| na | 3.285 | na | 3.184 | na | 3.244 |
| ti | 2.590 | AF | 2.394 | AF | 2.415 |
| AF | 2.349 | ti | 2.251 | ti | 2.233 |
| ri | 2.201 | pK | 2.054 | li | 1.894 |
| li | 2.173 | t | 2.048 | t | 1.889 |
| ... | | ... | | ... | |

Distribution of the Word Endings in the

| Hypothesis of the Hybrid System with word LM | | Human Reference Translation | | Training Data | |
|---|---|---|---|---|---|
| i | 35.961 | i | 30.402 | i | 30.496 |
| a | 15.117 | a | 16.925 | a | 16.868 |
| u | 7.958 | u | 10.333 | u | 10.320 |
| y | 4.906 | y | 6.890 | y | 7.122 |
| A | 4.459 | A | 6.432 | A | 6.252 |
| K | 3.285 | K | 5.520 | K | 5.249 |
| n | 2.590 | n | 4.956 | n | 4.716 |
| Y | 2.349 | Y | 4.436 | Y | 4.398 |
| F | 2.201 | F | 3.519 | F | 3.527 |
| t | 2.173 | t | 2.048 | t | 1.889 |
| ... | | ... | | ... | |

# Conclusion and Future Work

## Conclusion

- Word endings depend on the grammatical role of the word within the sentence. This leads to long-range dependencies, which are not well captured by the current models.

## Future Work

- Explore which features are useful to reduce errors in the word endings
- Find out whether the integration of the proposed diacritization features enhances the Arabic-English or English-Arabic translation systems

**Thanks for your interest!**

# References


T. Buckwalter. 2004. Arabic Morphological Analyzer version 2.0. LDC2004L02.

Mona Diab, Mahmoud Ghoneim, and Nizar Habash. 2007. Arabic Diacritization in the Context of Statistical Machine translation. In *Proceedings of the MT-Summit*, Copenhagen, Denmark.

Yousif A. El-Imam. 2004. Phonetization of Arabic: Rules and Algorithms. *Computer Speech and Language*, 18(4).

Tarek A. El-Sadany and Mohamed A. Hashish. 1989. An Arabic Morphological System. *IBM Systems Journal*, 28(4).

Ossama Emam and Volker Fischer. 2005. Hierarchical Approach for the Statistical Vowelization of Arabic Text. Technical report, IBM Corporation Intellectual Property Law, Austin, TX, US.

Ya'akov Gal. 2002. An HMM Approach to Vowel Restoration in Arabic and Hebrew. In *Proceedings of the ACL-02 Workshop on Computational Approaches to Semitic Languages*, Philadelphia, PA, USA.

Nizar Habash and Owen Rambow. 2007. Arabic Diacritization through Full Morphological Tagging. In *Proceedings of NAACL/HLT 2007. Companion Volume, Short Papers*, Rochester, New York, April.

Philipp Koehn, Hieu Hoang, Alexandra Birch an Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar ad Alexandra Constantin, and Evan Herbst. 2007. Moses: Open Source Toolkit for Statistical Machine Translation. In *Annual Meeting of ACL, demonstration session*, Prag, Czech Republic, June.

John Lafferty, Andrew McCallum, and Fernando Pereira. 2001. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In *Proceedings of the 18th ICML*.

Mohamed Maamouri, Ann Bies, and Seth Kulick. 2006. Diacritization: A Challenge to Arabic Treebank Annotation and Parsing. In *Proceedings of the British Computer Society Arabic NLP/MT Conference*.

Rada Mihalcea. 2002. Diacritics Restoration: Learning from Letters versus Learning from Words. In *Proceedings of the 3rd CICLing*, London, UK.

Husni Al-Muhtaseb Mustafa Elshafei and Mansour Alghamdi. 2006. Statistical Methods for Automatic Diacritization of Arabic Text. In *Proceedings of the Saudi 18th National Computer Conference (NCC18)*, Riyadh, Saudi Arabia, March.

Rani Nelken and Stuart M. Shieber. 2005. Arabic Diacritization Using Weighted Finite-State Transducers. In *Proceedings of the ACL 2005 Workshop On Computational Approaches To Semitic Languages*, Ann Arbor, Michigan, USA.

Franz Josef Och and Hermann Ney. 2003. A Systematic Comparison of Various Statistical Alignment Models. *Computational Linguistics*, 29(1):19–51.

# References


Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th ACL*, Philadelphia.

Michel Simard, Nicola Ueffing, Pierre Isabelle, and Roland Kuhn. 2007. Rule-Based Translation with Statistical Phrase-Based Post-Editing. In *Proceedings of the Second Workshop on Statistical Machine Translation*, Prague, Czech Republic, June.

Andreas Stolcke. 2002. SRILM – an Extensible Language Modeling Toolkit. In *International Conference on Spoken Language Processing*, Denver, USA.

Kristina Toutanova and Christopher D. Manning. 2000. Enriching the Knowledge Sources Used in a Maximum Entropy Part-of-Speech Tagger. In *Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora (EMNLP/VLC-2000)*, Hong Kong.

Dan Tufis and Adrian Chitu. 1999. Automatic Diacritics Insertion in Romanian Texts. In *Proceedings of COMPLEX'99 International Conference on Computational Lexicography*, Pecs, Hungary, June.

Dimitra Vergyri and Katrin Kirchhoff. 2004. Automatic Diacritization of Arabic for Acoustic Modeling in Speech Recognition. In *COLING 2004 Computational Approaches to Arabic Script-based Languages*, Geneva, Switzerland, August 28th.

Stephan Vogel, Ying Zhang, Fei Huang, Alicia Tribble, Ashish Venugopal, Bing Zhao, and Alex Waibel. 2003. The CMU Statistical Machine Translation System. In *Proceedings of MT-Summit IX*, New Orleans, Louisiana, USA, September.

Ying Zhang. 2006. SALM: Suffix Array and its Applications in Empirical Language Processing. Technical Report CMU-LTI-06-010, LTI, Carnegie Mellon University, Pittsburgh PA, USA.

Imed Zitouni, Jeffrey S. Sorensen, and Ruhi Sarikaya. 2006. Maximum Entropy Based Restoration of Arabic Diacritics. In *Proceedings of the 44th Annual Meeting of the ACL*, Sydney, Australia, July.