Language Model Based Arabic Word Segmentation

COLLEGE OF COMPUTER SCIENCE AND SOFTWARE ENGINEERING
DEPARTMENT OF INFORMATION AND COMPUTER SCIENECE

ICS 482, Natural Language Processing
Term 062

KING FAHD UNIVERSITY OF PETROLEUM AND MINERALS

NLP Presentation

Prepared for:
Mr. Husni Al-Muhtaseb

By:
Saleh Fahad Al-Zaid , ID: 222852

28 May 2007

Summary

The presentation is to introduce an algorithm to segment and acquire new Arabic stems form un-segmented Arabic corpus. It first describes segmentation using the pattern: prefix*-stem-suffix*, where a prefix is indicated by # character and a suffix by + character.

The process of the algorithm includes:

1- Morpheme Segmentation
This uses trigram language model, that is using p(mi | mi-1, mi-2) applies first to manually segmented corpus. After that, the trigram probability table and "prefixes and suffixes" table are built for each segment of the manually segmented corpus.

Next step is to build a decoder for morpheme segmentation to find the morpheme sequence in new un-segmented corpus which maximizes the trigram probability of the input sentence, as in the formula:

SEGMENTATIONbest = *Argmax* IIi=1,N p(*mi*|*mi-1 mi-2*)
N = number of morphemes in the input

Using this formula, we try to find the possible segmentation of a new word. This includes the following steps:

i) Identify all of the matching prefixes and suffixes from the "prefixes and suffixes" table,
ii) Further segment each matching prefix/suffix at each character position, and
iii) Enumerate all *prefix*-stem-suffix** sequences derivable from (i) and (ii).

2- New Stems Acquisition
After that starts the process of acquiring new stems from new large un-segmented corpus. This process includes the following:

Initialization: Develop the seed segmenter Segmenter0 trained on the manually segmented corpus Corpus0, using the language model vocabulary, Vocab0, acquired from Corpus0.

- **Iteration:** For i = 1 to N, N = the number of partitions of the unsegmented corpus:
    - i. Use Segmenteri-1 to segment Corpusi.
    - ii. Acquire new stems from the newly segmented Corpusi. Add the new stems to Vocabi-1, creating an expanded vocabulary Vocabi.
    - iii. Develop Segmenteri trained on Corpus0 Corpusi with Vocabi.

Finally, the performance evaluation of this algorithm on a given corpus described by the formula:
E = (*number of incorrectly segmented tokens / total number of tokens*) *x 100.*

List of References

Paper: Language Model Based Arabic Word Segmentation, 2003. By: Young-Suk Lee, Kishore Papineni, Salim Roukos, Ossama Emam and Hany Hassan.

Obstacles Faced

- Try to understand new terminologies like segmentation what does it mean.
- Understanding paper's algorithms and the idea behind them.

Things Learned and Skills Gained

- Comprehend that Arabic Natural Processing is a hot topic and has many papers published and effort done on it.
- Segmentation is a good idea to understand how we can get a stem (and maybe further a root) from a word by dividing it into determined patterns.
- Experience with Buckwalter's form of Arabic text.
- Language model based Arabic word segmentation gave me the experience of an application implementing the Tri-Gram model.

Recommendation

Increase focusing on NLP applications like language model based Arabic word segmentation paper during the course. Because it helps the students for more understanding of course topics.

True/ False Questions

1. Language model based Arabic word segmentation uses bi-gram model.
   Answer: False, *Correct answer*: it uses tri-gram model

2. The output of morpheme segmentation phase are "prefixes and suffixes" and the trigram probability tables.
   Answer: True.

3. The performance evaluation of language model based Arabic word segmentation algorithm can be calculated by:

E = (*number of correctly segmented tokens / total number of tokens*) *x 100*

Answer: False, *Correct answer*:
E = (*number of incorrectly segmented tokens / total number of tokens*) *x 100*