HYBRID WORD/PART-OF-ARABIC-WORD LANGUAGE MODELS FOR ARABIC TEXT DOCUMENT RECOGNITIION

Mohamed Faouzi BenZeghiba, Jérôme Louradour and Christopher Kermorvant

ABSTRACT

This paper describes a simple approach to generate an efficient hybrid word/Part-of-Arabic-Word (PAW) Language Model (LM). Less frequent words are decomposed into PAWs, which are then with the most frequent words to generate a hybrid flat language model. Evaluation experiments are conducted under three different tasks (Maurdor printed/handwritten and Khatt). Hybrid LMs, systematically, outperform word LMs, Moreover, they require less memory.

MOTIVATION

Issues:

- Language model is a main component in State-of-the-art Arabic text recognition systems.
- Issue: Detection and the recognition of Out-Of-Vocabulary words (OOV)
- Increasing the vocabulary might increase the confusability.

Some solutions:

- Use of sub-word based LMs.
- Question: What is the best sub-word unit?
  - Character LM: higher order character LMs might outperform word LMs.
  - Longer sub-word unit to capture a wider context is more appropriate.
- Our investigation: Use of Part-of-Arabic-Word (PAW).
- Widely used in Arabic natural language processing applications.

PAW decomposition:

- Arabic script is cursive.
- Exceptions: Six Arabic characters (ا, د, ذ, ر, ز, و), cannot be connected to the character on the left.
- Words containing one or more of these six characters can be decomposed into PAWs.
- Example: The word المَدرَسَة is decomposed into four PAWs: ا, لمَد, رَ and سَة.

THE PROPOSED HYBRID LM

LM generation:

- Text normalization and tokenization: including space tokenization
- Word vocabulary extraction and word LM generation.
- PAW decomposition: applied to less frequent words in the word vocabulary
- Hybrid vocabulary generation: Integrating the most frequent words and the resulted PAWs.
- Hybrid flat n-gram LM generation

Example:

(a) مَعرِفَة النّص العرَبِي المَخطوط

(b) معرفة ـ | النص | ـ | العربي | ـ | المخطوط

(c) معر | فة | ـ | ا | لنص | ـ | ا | لعر | بي | ـ | ا | لمخطو | ط

Advantages:

- Simple decomposition rule.
- Reduction in vocabulary size.
- Reduction in OOV rate.
- Better estimation of the n-gram statistics.
- Results: Better prediction capabilities of the LM.

Post-processing:

- Removing the normal space (to connect PAWs).
- Replacing the space token with normal space.

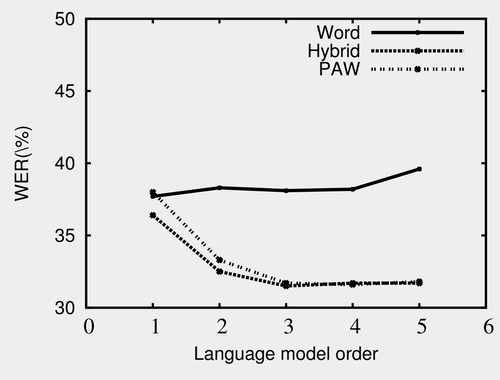
SYSTEM DESCRIPTION

- Processing: Text line detection for Maurdor data.
- Optical models: Multi-Directional LSTM Recurrent Neural Network
- Dropout for regularization.
- LM models: Different statistical LM types with diffrent orders.
- Decoding: Performed with Kaldi toolkit.

EXPERIMENTAL RESULTS

KHATT database:

- Effect of the LM order

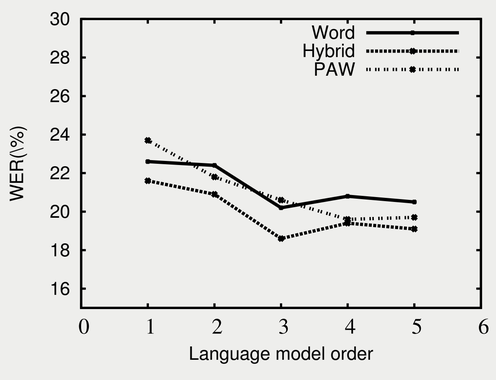


- Final results

Word Error Rate				
LM data		LM type		
	Dataset	Word	Hybrid	PAW
Train	Dev	**37.7**	**31.5**	**31.6**
	Test	**40.2**	**33.1**	**33.0**
Train+Dev	Test	**37.8**	**31.3**	**30.9**
OOV rate				
Train+Dev	Test	**24.9**	**9.1**	**9.1**

Maurdor printed

- Effect of the LM order

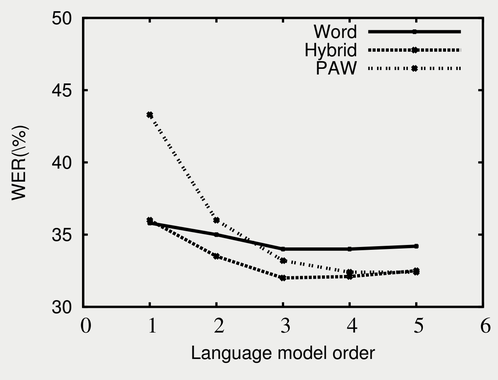


- Final results

Word Error Rate				
LM data		LM type		
	Dataset	Word	Hybrid	PAW
	Dev	**20.2**	**18.6**	**19.6**
Train	**Test**	**26.6**	**22.2**	**23.5**
Train+Dev	**Test**	**26.5**	**22.2**	**23.1**
OOV rate				
Train+Dev	**Test**	**15.4**	**8.4**	**7.9**

Maurdor Handwritten

- Effect of the LM order



- Final results

Word Error Rate (%)				
LM data		LM type		
	Dataset	Word	Hybrid	PAW
	Dev	**34.0**	**32.0**	**32.4**
Train	**Test**	**35.1**	**33.4**	**34.5**
Train+Dev	**Test**	**34.8**	**33.2**	**33.5**
OOV rate (%)				
Train+Dev	**Test**	**15.1**	**10.8**	**10.4**

PERSPECTIVES

- Hybrid LM for large vocabulary task.
- Hierarchical LMs incorporating words, PAWs and characters.
- Combination of several word decomposition approaches.

{mfb,jl}@a2ia.com, kermorvant@teklia.com