

# Recent Developments of the Cambridge Arabic Speech-to-Text Systems

F. Diehl, M.J.F. Gales, X. Liu, J. Park, M. Tomalin & P.C. Woodland

9 July, 2010



Cambridge University Engineering Department

# Contents

- Introduction to the GALE project
- LVCSR system overview
- Arabic LVCSR – what is the challenge?
- Morphological decomposition using MADA
- Graphemic and phonetic modelling
- Neural network language model



# Introduction to the GALE Project

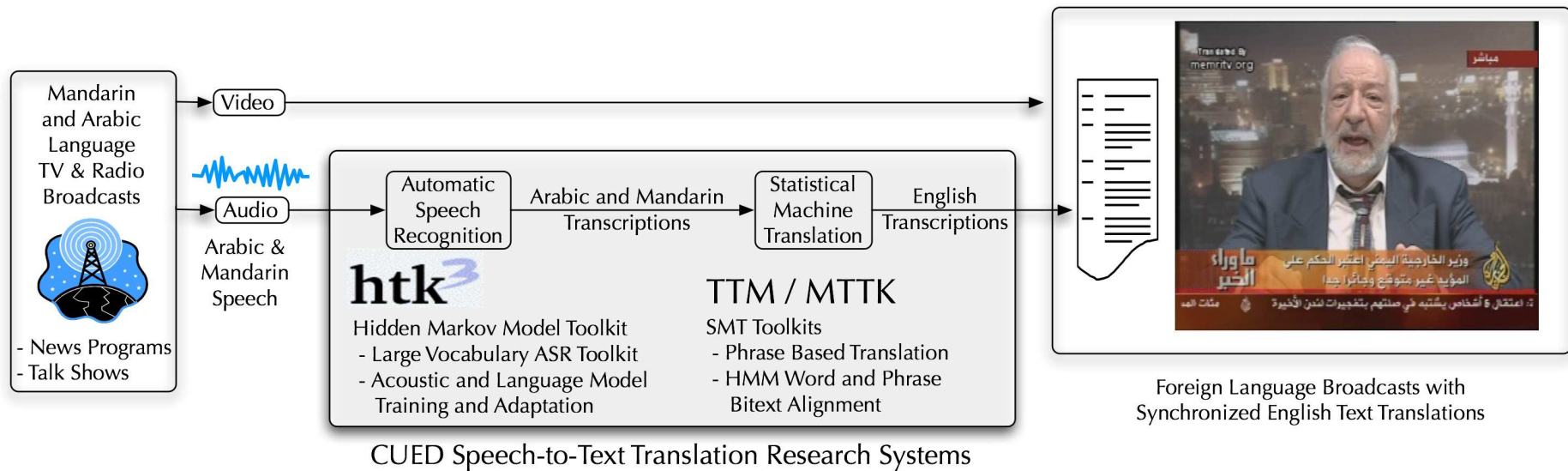


## Global Autonomous Language Exploitation (GALE)

- **DARPA** funded research program (Thanks to DARPA!)  
(<http://www.darpa.mil/ipto/programs/gale/gale.asp>)
- Goal:
  - ▷ "...develop and apply computer software technologies to absorb, **analyse** and interpret **huge volumes of speech** and text in **multiple languages... delivering** pertinent, consolidated **information...in easy-to-understand forms to...monolingual English-speaking analysts...**"
- GALE consists of three major engines:
  - ▷ **Transcription**
  - ▷ Translation
  - ▷ Distillation



# Global Autonomous Language Exploitation (GALE)



## CUED within the GALE Project

- Two competing groups: AGILE, ROSETTA
- The **AGILE** team: BBN, LIMSI, ..., and **CUED**
- The objective of AGILE consists in:
  - Transcription of Arabic/Chinese broadcast news/conversation data
  - Translation of the transcribed data to English
- Final STT result: **system combination** CUED, LIMSI, and BBN

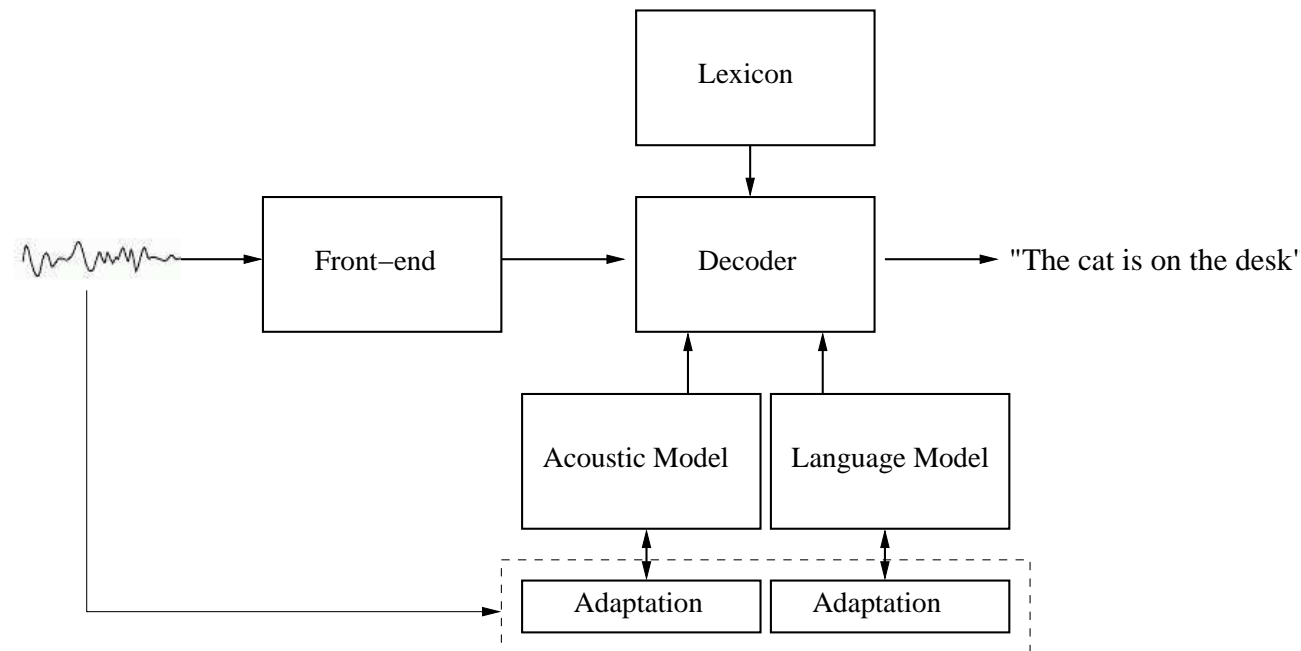


## LVCSR System Overview



## LVCSR System Overview

- What is a LVCSR system?



- approx. 1350 hrs acoustic training data
- 9k states, 324k mixtures
- approx. 1G words LM training data
- 4-gram LMs



## LVCSR System Overview

- Broadcast News/Conversational transcription tasks
  - Single audio stream with many talkers, styles, noise conditions, bandwidths
  - Need to segment for normalisation/adaptation
  - Vocabulary changes with news stories (BN) or casual conversation style (BC)
- Front-end parametrisation
  - Basic front-end uses cepstral parameters (typically 12 cepstra + energy/c0)
  - Perceptual Linear Prediction (PLP) is widely used.
- Lexicon design & Acoustic models
  - Standard phone model is widely used
  - HMM based stochastic modelling of speech distributions



- Language models

- gives probabilities of sentences
- N-gram models so that the probability of a word string  $w$  is

$$\begin{aligned} P(w) &= P(w_1, w_2, w_3, \dots, w_n) \\ &= P(w_1) P(w_2|w_1) P(w_3|w_1, w_2), \dots, P(w_n|w_1, \dots, w_{n-1}) \\ &\approx \prod_{k=1}^T P(w_k|w_{k-1}, \dots, w_{k-N+1}) \end{aligned}$$

- model size grows exponentially with the dictionary size
- many unseen model parameters
- use back-off strategy for unseen N-grams

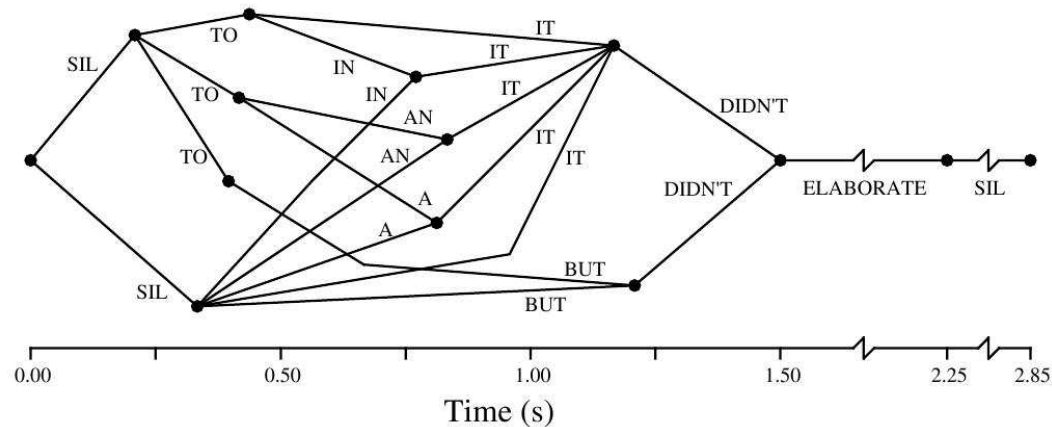


## LVCSR System Overview

- Large vocabulary decoder
  - Find 1-best or N-best / lattice of recognition alternatives

$$\hat{w} = \max_w P(o|w)P(w)$$

- Word lattices

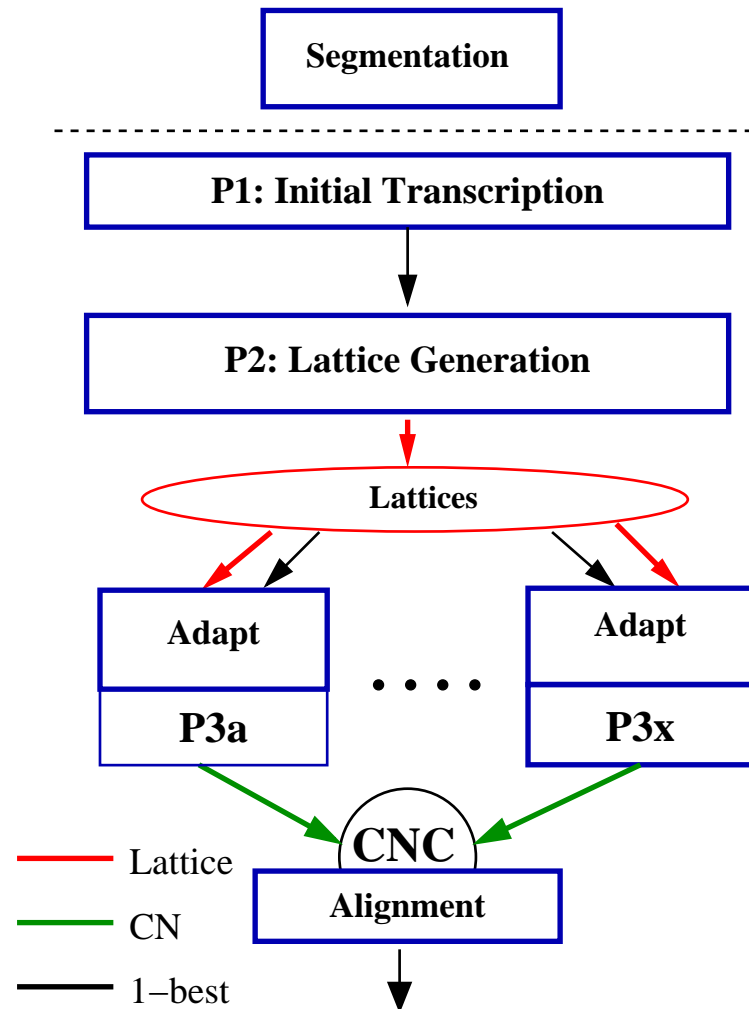


- ⇒ A set of nodes that correspond to points in time
- ⇒ A set of arcs encoding word-word transitions (acoustic/language scores)



## LVCSR System Overview

- Decoding Strategy

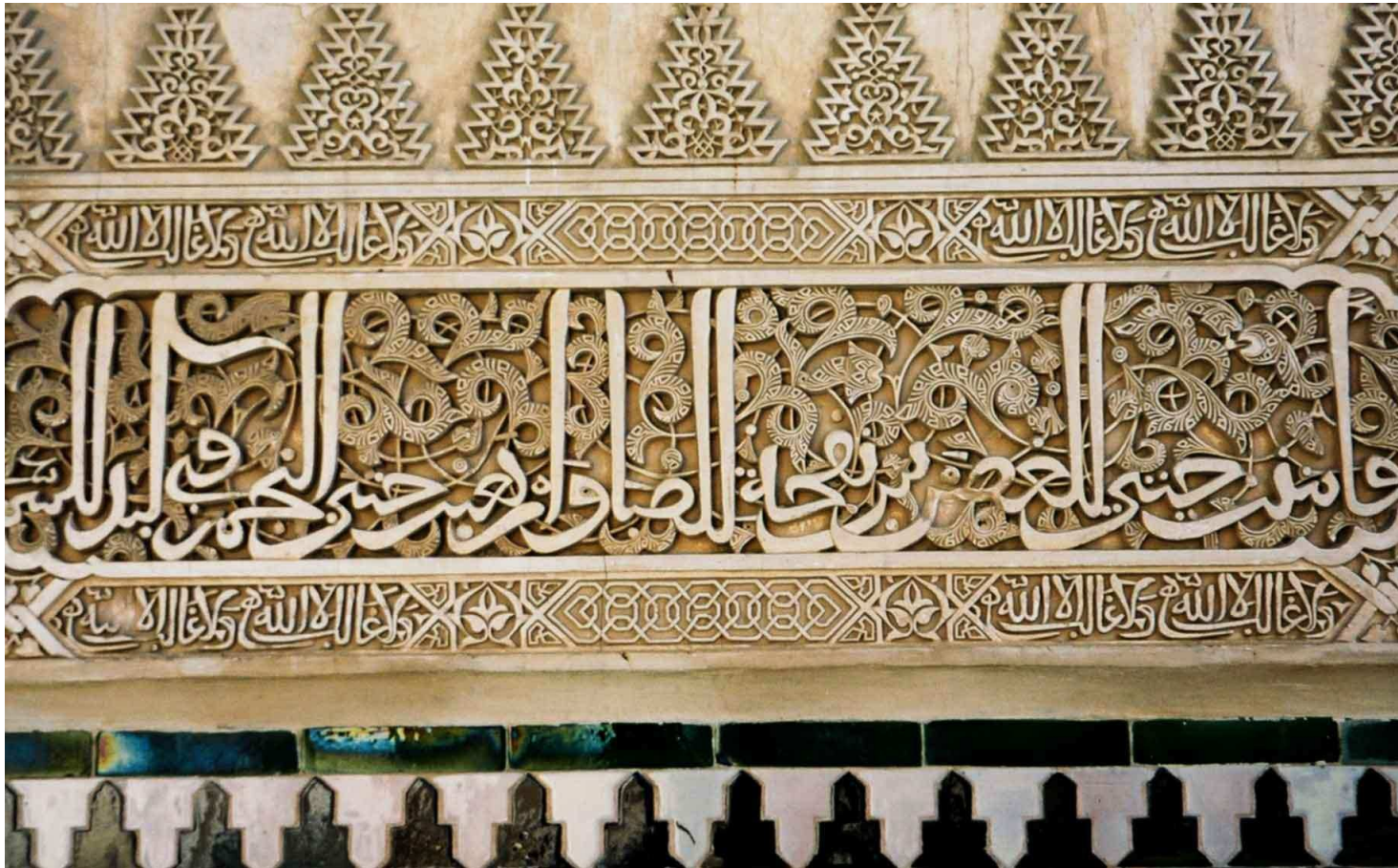


- Multi-pass combination framework

- P1 used to generate **initial** hypothesis
- P1 hypothesis used for rapid **adaptation**
  - \* LSLR, diagonal variance transforms
- P2 lattices generated for rescoring
  - \* **apply complex LMs**
- P3 adaptation
  - \* **acoustic model rescoring**
  - \* 1-best CMLLR
  - \* Lattice-based MLLR
  - \* Lattice-based full variance
- CN decoding / combination



## Arabic LVCSR – what is the challenge?



## Arabic LVCSR – what is the challenge?

- Issues with Arabic STT:

- Arabic is a morphologically complex language
  - ⇒ high Out-Of-Vocabulary (OOV) rates
  - ⇒ sparse LM training data
- Modern Standard Arabic (MSA) is written without short vowels
  - ⇒ difficult dictionary generation
- There are many different Arabic dialects

- Proposed solutions:

- ⇒ morphological decomposition
- ⇒ apply graphemic and phonetic models in combination
- ⇒ neural network LMs



## Arabic LVCSR – what is the challenge?

والاعتراف بالخطا فضيلة

wAlAEtrAf bAlxTA fDylp

wAlAiEtirAfa bAlxaTaA faDiylapF



## Morphological Decomposition



## Morphological Decomposition

- Arabic is a highly inflected agglutinative language
  - many lexical variants per root word → high OOV rates  
(e.g. 59k English dict: ~ 0.5% OOV, 350k Arabic dict: ~ 1.0% OOV)
  - sparse language model training material
- MADA (Habash'06) morphological decomposition alleviates these problems by
  - splitting prefixes from stems:  $IAIREb \rightarrow I+ AIREb$
  - stem normalisation:  $AlAyrAnY \rightarrow AlMyrAnY$
- MADA 'D2 DIAC' decomposition scheme used:
  - separates five proclitics ( $l, b, k, w, f$ ) from associated word roots
  - provides stem normalisation
  - provides vowelisation (diacritization)



## Morphological Decomposition

- Stem normalisation makes the morpheme-to-word conversion a non-trivial task:
  - no simple prefix-stem gluing possible  $\rightarrow$  morpheme-to-word mapping
  - MADA is context sensitive  $\rightarrow$  a unigram mapping is suboptimal
- Morpheme-to-word mapping viewed as Machine Translation (MT) task:
  - translate the morpheme sequence ( $S$ ) to the word sequences ( $T$ )
  - linear alignment ( $T, S$ ) between source and target token sequences exists

$\Rightarrow$  N-gram SMT approach

$$p(T, S) = \prod_{k=1}^K p((t, s)_k | (t, s)_{k-1}, \dots, (t, s)_{k-N+1})$$

- Decoding: likelihood maximisation of  $p(T|S)$  w.r.t..  $T = (t_1, t_2, \dots, t_K)$ ,



# Graphemic and Phonetic Modelling



## Graphemic and Phonetic Modelling

- Dictionary generation:
  - Arabic is commonly written without diacritic markers indicating short vowels
  - Many phonetic word forms for one graphemic word form

Graphemic Form	Vowelised Form	Meaning
<i>ktb</i>	<i>kataba</i> <i>kutiba</i> <i>kitab</i> <i>kutubun</i> <i>kutubu</i>	he writes it is written/fated/destined book (indefinite) books (indefinite) books (definite)

- Use of graphemic and phonetic systems in combination
  - ⇒ graphemic system: good coverage and robustness
  - ⇒ phonetic system: better accuracy but some words not available



## OOV Rate Reduction by MADA and Gra/Pho Modelling

- Pronunciation generation:
  - Word-based system: pronunciations provided by ‘Buckwalter’
  - MADA-based system: pronunciations provided by MADA

Wordlist		Testset (OOV)	
Type	Size	eval07	dev08
Word-phonetic	260k	3.39	2.03
Word-graphemic	350k	1.26	1.14
MADA-phonetic	171k	1.33	1.38
MADA-graphemic	331k	0.54	0.63

⇒ MADA: OOV rate reduction by nearly a factor of 2

⇒ Improved coverage of graphemic vs phonetic systems



## WER Reduction by MADA and Gra/Pho Modelling

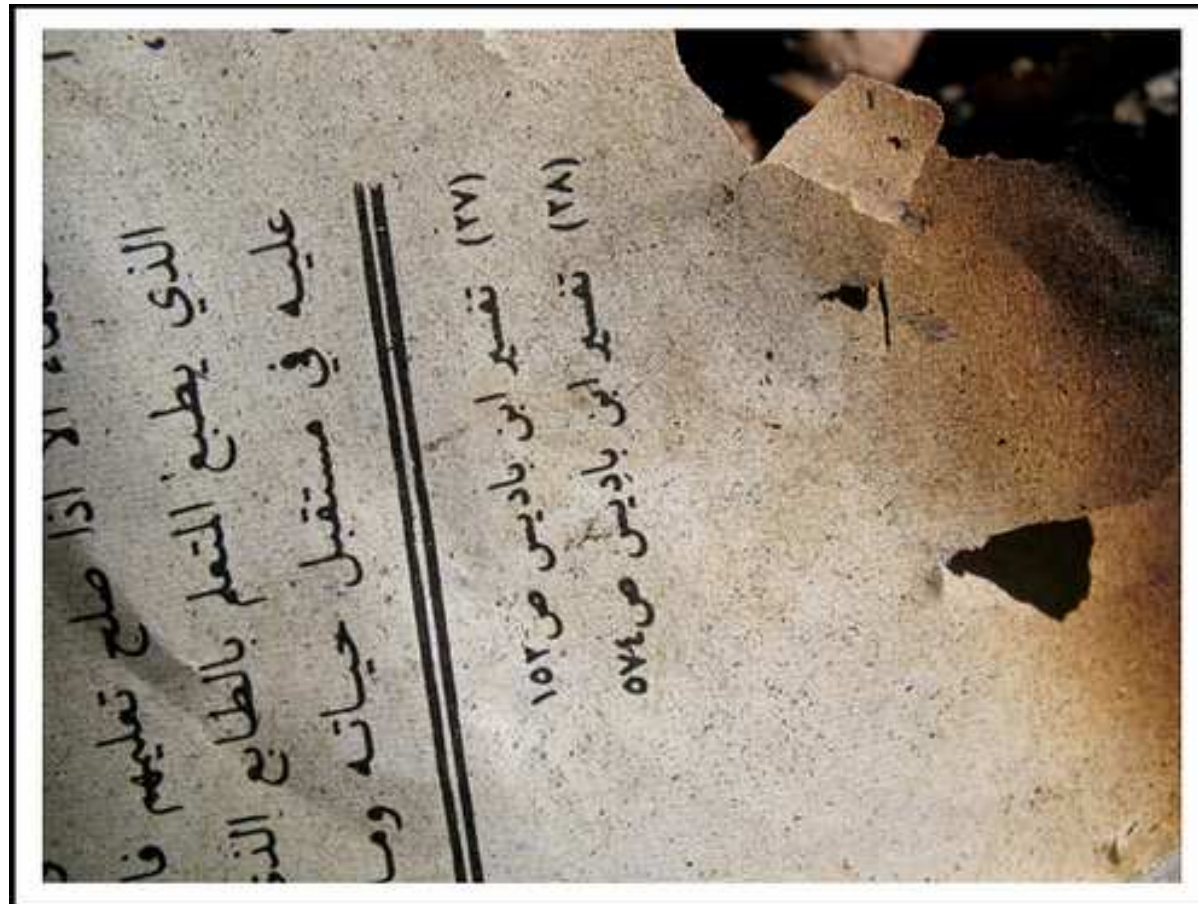
- Comparison of graphemic/phonetic word/morpheme systems:
  - ‘G’ = graphemic system
  - ‘V’ = phonetic system (**V**owelised system)
  - ‘ $M_a$ ’ = MADA morphological decomposition

System	Configuration	Testset (WER)	
	mada	eval07	dev08
<b>G</b>	-	14.1	14.9
<b>G</b> $M_a$	✓	13.6	14.3
<b>V</b>	-	12.9	14.0
<b>V</b> $M_a$	✓	12.4	13.5
<b>V</b> $M_a \oplus$ <b>G</b> $M_a$	✓	12.0	12.9

⇒ MADA provides reductions of the WER by typically 0.5% absolute



## Neural Network Language Model



## Neural Network Language Model

- Problems of standard N-gram LMs:

$$P(w) = \prod_{k=1}^T P(w_k | w_{k-1}, \dots, w_{k-(N-1)})$$

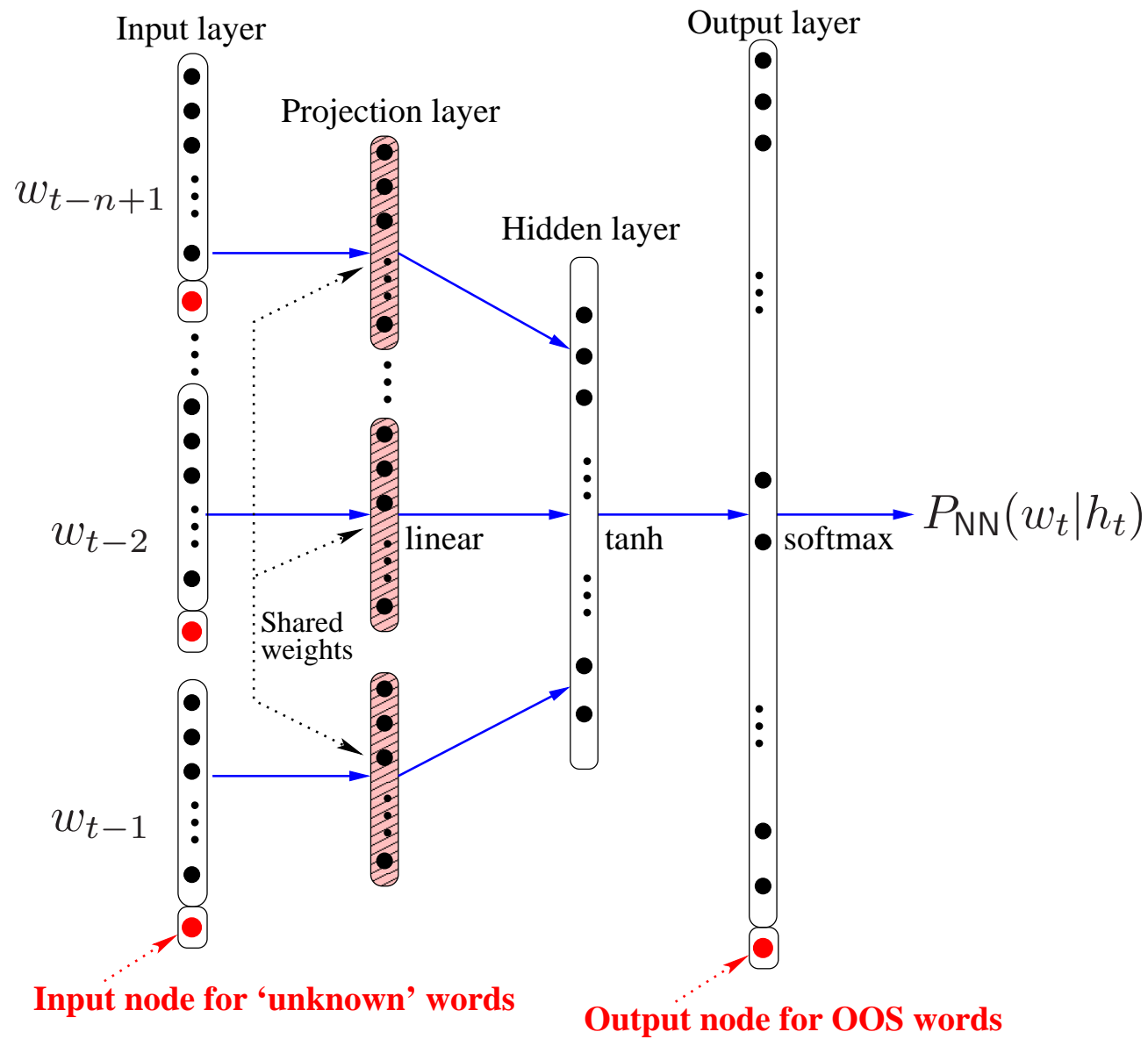
- **unstructured model**  $\Rightarrow$  direct estimation (counting) of N-gram probabilities
  - **no interpolation** of unseen model parameters  $\Rightarrow$  back-off strategies
  - **difficult LM adaptation**, many parameters little data
- Neural network LMs offer:
    - **a structured modelling** approach with a continuous feature space
    - **parameter interpolation** possible  $\Rightarrow$  no back-off strategy needed
    - **cope better** with the **data sparsity** problem of standard N-gram LMs
    - provides a way for **LM adaptation**



## Neural Network Language Model

- General description of NNLM
  - Projects a set of contexts  $h_t = w_{t-1} \dots w_{t-n+1}$  onto a continuous space
  - Calculates the LM probability for each word given a history,  $P(w_t = i | h_t)$
- Architecture of network with OOS node
  - Fully-connected multi-layer perceptron (MLP) structure
  - The inputs to the network ( $N - 1 * 100k$ )
    - ⇒ indices of the  $n - 1$  history words in the input vocabulary  $V_{in}$
  - Between input and output layers ( $100 \times 400$ )
    - ⇒ two hidden layers for projection and non-linear probability estimation
  - The outputs of the network ( $20k$ )
    - ⇒ posterior probabilities of all words following a given history





## Neural Network Language Model

- Interpolation with N-gram LM
  - Smaller dictionary coverage than N-gram LM  $\Rightarrow$  LM interpolation

$$P(w_t|h_t) = \lambda P_{\text{NG}}(w_t|h_t) + (1 - \lambda) \tilde{P}_{\text{NN}}(w_t|h_t)$$

- Probability normalisation

$$\tilde{P}_{\text{NN}}(w_t|h_t) = \begin{cases} P_{\text{NN}}(w_t|h_t) & w_t \in V_{\text{sl}} \\ \beta_S(w_t|h_t) P_{\text{NN}}(w_{\text{oos}}|h_t) & \text{otherwise} \end{cases}$$

$$\beta_S(w_t|h_t) = \frac{P_{\text{NG}}(w_t|h_t)}{\sum_{\tilde{w}_t \notin V_{\text{sl}}} P_{\text{NG}}(\tilde{w}_t|h_t)}$$

$\Rightarrow$  very expensive in decoding time

- Approximated normalization

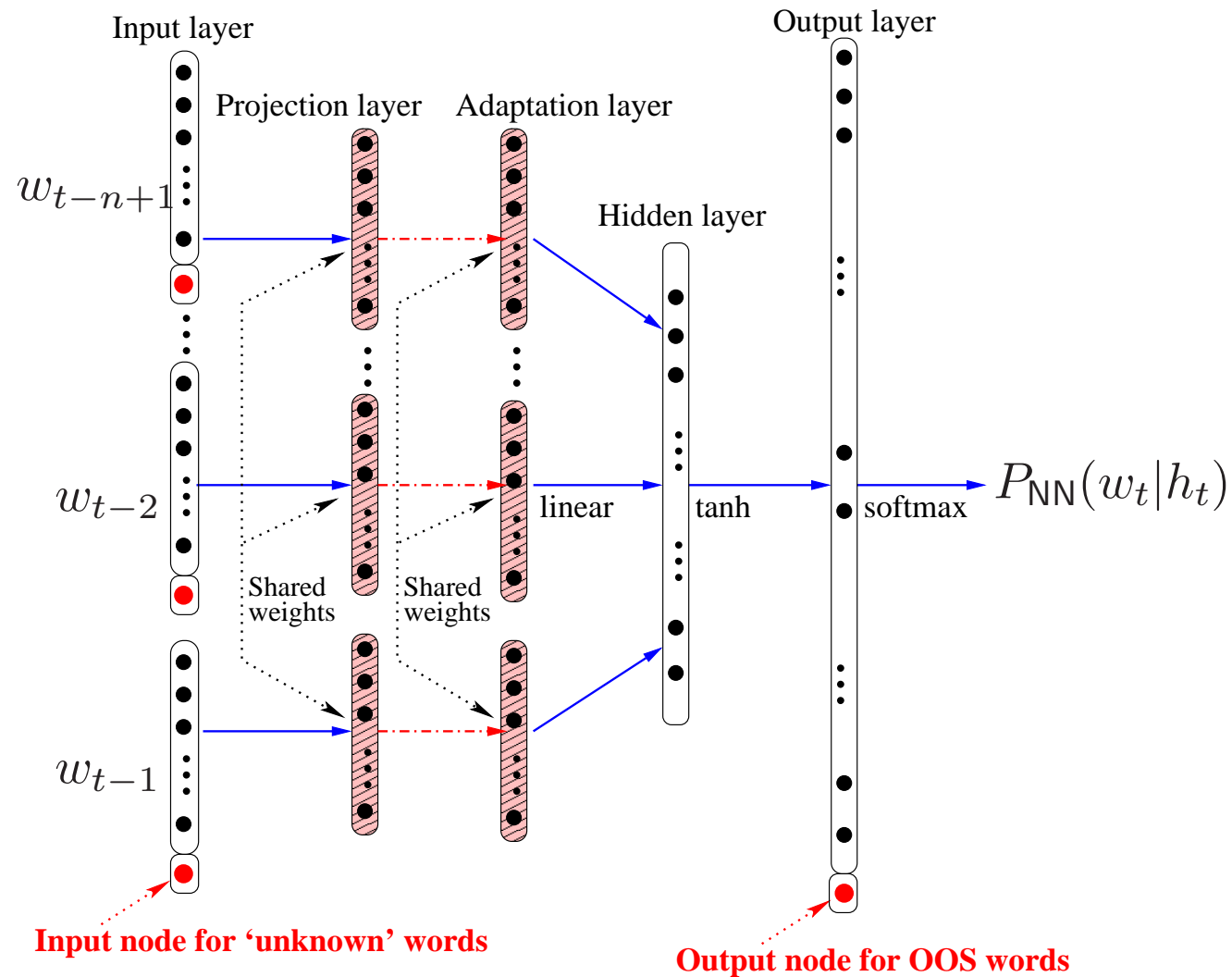
$$\tilde{P}_{\text{NN}}(w_t|h_t) = \begin{cases} P_{\text{NN}}(w_t|h_t) & w_t \in V_{\text{sl}} \\ P_{\text{NG}}(w_t|h_t) & \text{otherwise} \end{cases}$$



## Neural Network Language Model

- LM adaptation to a particular domain or task
  - Improve robustness to varying styles or tasks
  - Data sparsity issue
    - ⇒ impractical adaptation of N-gram prob. (limited amounts of data)
  - Continuous space representation in NNLM
    - ⇒ continuous feature space allows parameter interpolation
    - ⇒ stronger generalisation ability





## Neural Network Language Model

- System evaluation results

System	LM	WER	
		eval07	dev08
$V_{mlp}$	NG	12.4	13.5
	NG + NN.OOS	12.3	13.1
	NG + NN.OOS.adapt	12.2	13.1

- Single branch system
  - ⇒ consistent WER reductions of 0.2%-0.4% absolute over the N-gram LM
  - ⇒ further improvements after NNLM adaptation
- Similar trends are also found in case of system combination



## Conclusions

- **MADA morphological decomposition** significantly reduces the OOV problem
- Combined **graphemic/phonetic modelling** copes with the dictionary generation problem
- **NNLM**
  - ⇒ **cope with the data sparsity problem** of Arabic LM modelling
  - ⇒ provides a method for **LM adaptation**
- **Current research:**
  - ⇒ **improved forms of NNLM adaptation**
  - ⇒ **incorporation of syntactic features into a NNLM**



Thank you for your attention

