



# Broadcast Speech Transcription and Translation



A. de Gispert, X. A. Liu, W. J. Byrne, M. J. F. Gales & P. C. Woodland  
Cambridge University Engineering Department  
{ad465,xl207,wjb31,mjfg,pcw}@eng.cam.ac.uk

## 1 INTRODUCTION

- **Key research areas for speech translation are:**
  - Speech Recognition (ASR) / audio Transcription.
  - Statistical Machine Translation (SMT).
  - Integration between ASR and SMT.
- **System development involves:**
  - Developing statistical models for speech and language tasks.
  - Parameter estimation from large data sets.
  - Finding the best hypothesis / speech segment etc.

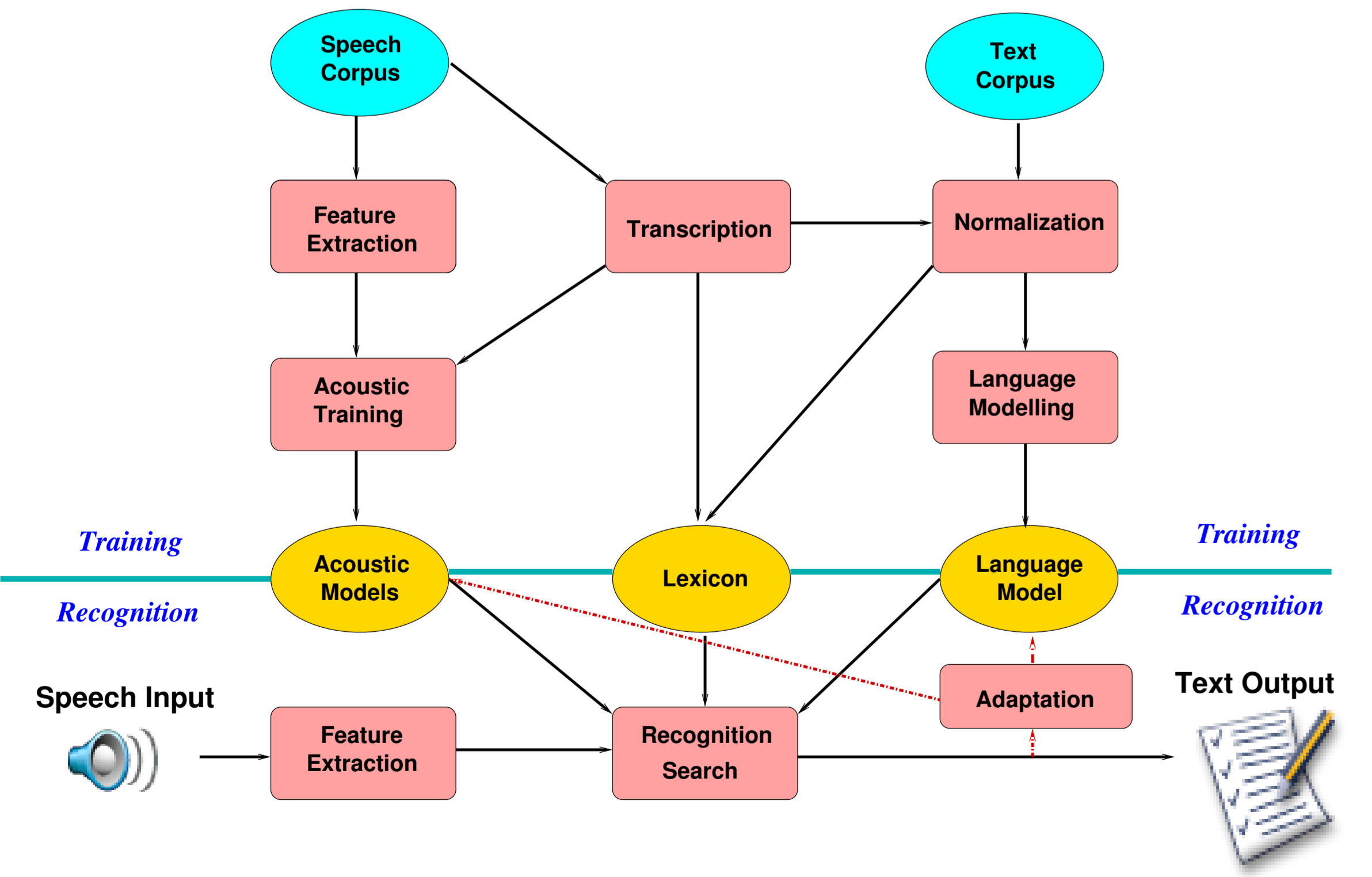
• **Both ASR and SMT can be formulated using a Source-Channel model:**

<u>Transcription</u>	<u>Translation</u>
Input - an utterance $A$	Input - a foreign sentence $F$
Output - a transcription $\hat{W}$	Output - an English sentence $\hat{E}$
$\hat{W} = \text{argmax}_W P(W A)$	$\hat{E} = \text{argmax}_E P(E F)$
$= \text{argmax}_W \frac{P(A W)P(W)}{P(A)}$	$= \text{argmax}_E \frac{P(F E)P(E)}{P(F)}$
$= \text{argmax}_W \underbrace{P(A W)}_{\text{Acoustic Model}} \underbrace{P(W)}_{\text{Source Language Model}}$	$= \text{argmax}_E \underbrace{P(F E)}_{\text{Translation Model}} \underbrace{P(E)}_{\text{Source Language Model}}$

Both use Maximum A Posteriori search using models learned from data.

## 2 SPEECH TRANSCRIPTION

• **Classic ASR system architecture:**

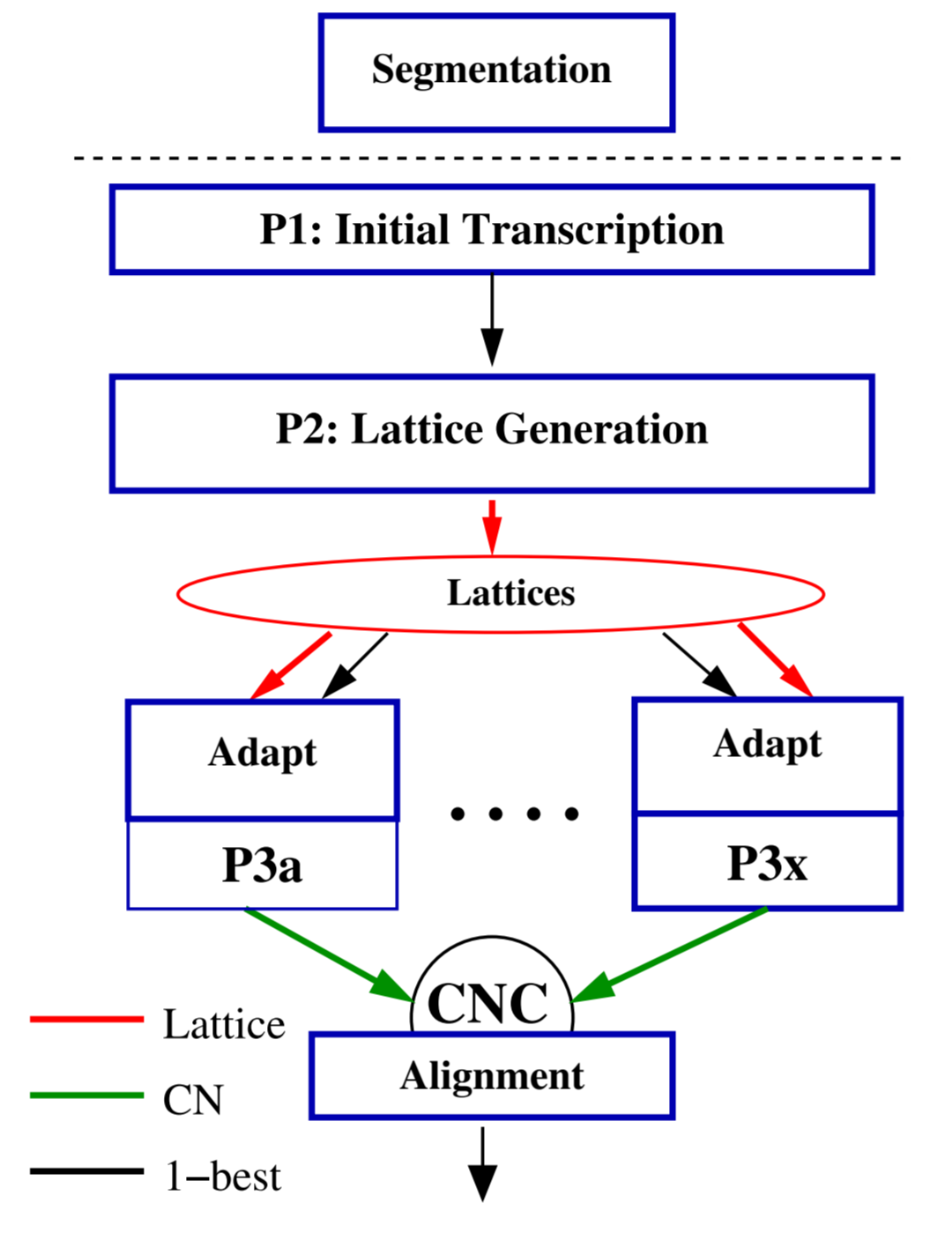


- **Statistical models typically trained using:**
  - more than 1500 hours of speech

- more than 2500 millions of words of text
- continuous density HMMs using cepstral features
- $N$ -gram language models
- maximum likelihood and discriminative training
- parameter tying and smoothing
- normalization and adaptation techniques
- combining outputs from complimentary systems

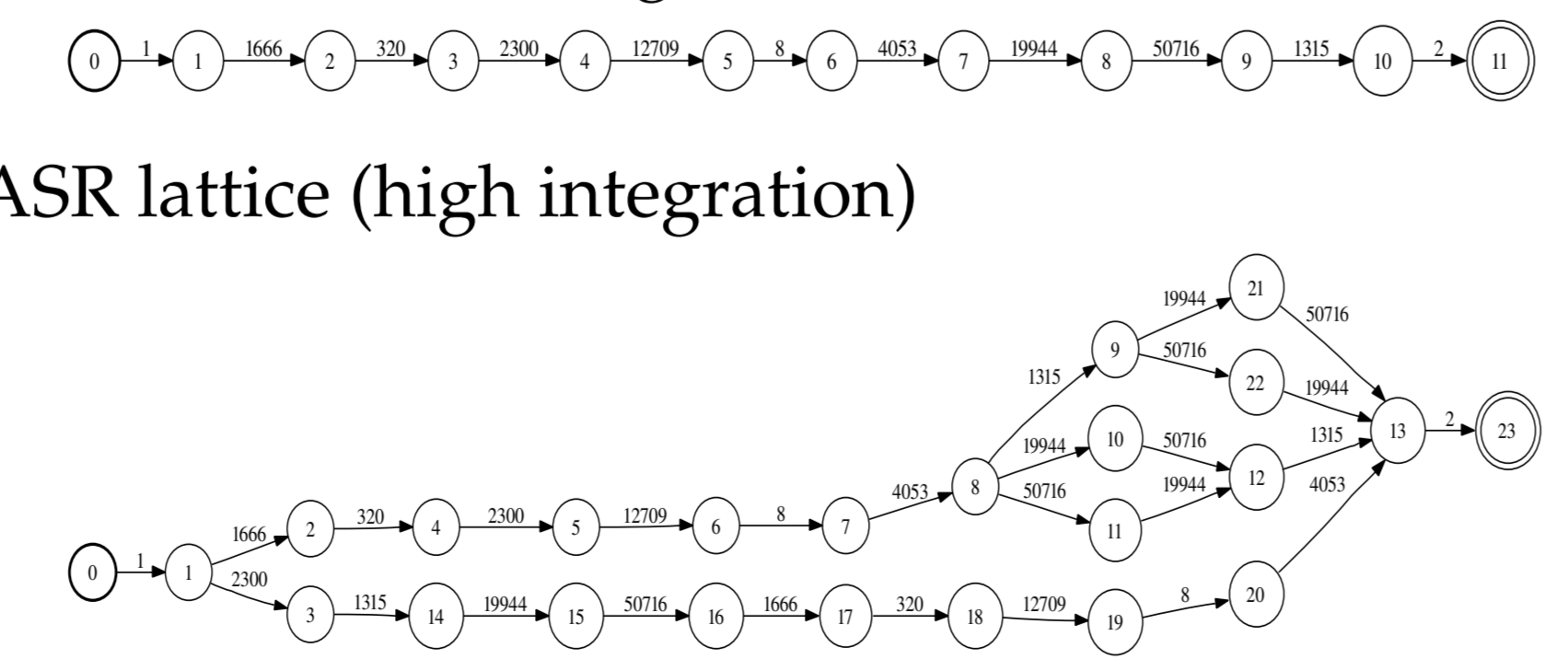
• **Broadcast speech transcription system:**

- Find speech and cluster into homogeneous regions
- Search for best possible word strings for each segment using acoustic and language models
- Multiple recognition passes: **lattices** constrain search space in later stages
- Unsupervised adaptation “tunes” models to speech
- Can run system faster if “prune” search more



## 3 ASR AND SMT INTEGRATION

- **Sentence boundary detection:**
  - SMT systems require well-defined sentential structures.
  - Post-processing ASR system outputs to locate sentence boundaries.
- **Original STT outputs:** 各位好欢迎您收看中文国际频道的今日关注今天呢我们要讨论的是伊朗核问题.
- **Post-processed STT outputs:** Sentence 1: 各位好欢迎您收看中文国际频道的今日关注. Sentence 2: 今天呢我们要讨论的是伊朗核问题.
- **Translating multiple recognition outputs:**
  - feeding ASR modelling information into translation.
  - flexible decoding framework for translation.
  - translate 1-best ASR (low integration)
  - translate ASR lattice (high integration)



## 4 GENERATIVE MODEL OF SPEECH TRANSLATION

- **Noisy channel model for speech translation**
  - Translation from target to source: search for most probable source sentence to have generated the target sentence

Target Speech	Target Sentence	Target Phrase	Source Phrase	Source Sentence
$A$	$\leftarrow t_1^J$	$\leftarrow v_1^R$	$\leftarrow u_1^K$	$\leftarrow s_1^I$
<b>Models</b>	$P(A t_1^J)$	$P(t_1^J v_1^R)$	$P(v_1^R u_1^K)$	$P(u_1^K s_1^I)$
<b>FSMs</b>	$\mathcal{L}$	$\Omega$	$\Phi$	$\mathcal{W}$
	ASR Word Lattice	Target Phrase Segmentation Transducer	Phrase Translation, Reordering Transducer	Source Phrase Segmentation Model Transducer

– The final translation is given by

$$\hat{s}_1^I = \text{argmax}_{s_1^I} \left\{ \max_{t_1^J, v_1^R, u_1^K, K} P(A, t_1^J, v_1^R, u_1^K, s_1^I) \right\}.$$

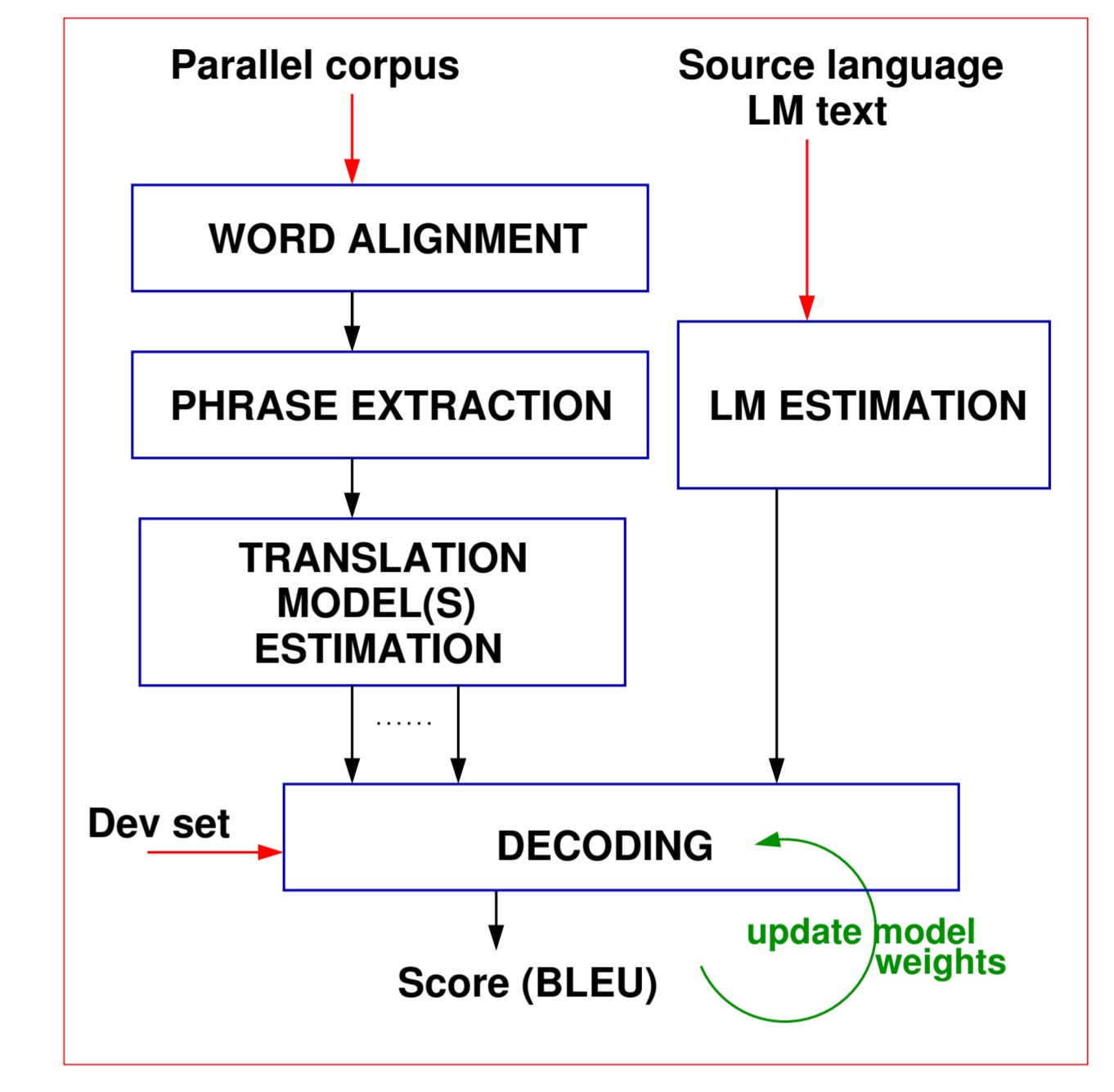
- **Implementation with Weighted Finite State Transducers**
    - Translation is performed using libraries of standard FST operations
    - Architectures may be limited, but easy to work on components
    - No special-purpose decoder required
    - Standard best-path search through the translation FST  $\mathcal{T}$
- $$\mathcal{T} = \mathcal{G} \circ \mathcal{W} \circ \Phi \circ \mathcal{Q}$$
- Efficient implementation with OpenFST ([www.openfst.org](http://www.openfst.org))

## 5 TRANSLATION SYSTEM TRAINING

- **Sentence-aligned parallel corpora**
  - Arabic $\leftrightarrow$ English  $\sim$  150M words (6M sent.)
  - Chinese $\leftrightarrow$ English  $\sim$  240M words (10M sent.)
- **English monolingual data**
  - Used for first-pass and rescoring LM

– Currently  $\sim$  5 billion words

• **Standard training stages:**



- **Minimum Error Training**
  - Find optimal models weights
  - Optimize evaluation metric (BLEU) on development set
- **Lattice Rescoring and MBR**
  - Rescoring LM: zero cut-off 5-gram
  - Phrase segmentation 2-gram
  - Model 1 lattice-to-string alignment
  - Minimum Bayes Risk decoding of final N-best lists

## 6 RESULTS

- **Text and speech translation performance**
  - BLEU: Automatic metric that measures similarity between hypothesis and a set of human-translated references
  - Dev/Test size:  $\sim$  60k words
  - Scores: 4 references for text, 1 reference for audio

Task	Condition	DEV SET	TEST SET
Arabic $\rightarrow$ English	clean text	54.02	53.70
Chinese $\rightarrow$ English	clean text	31.07	32.24
Chinese $\rightarrow$ English	audio reference	18.60	18.10
	ASR lbest	–	16.59

- ASR lattice translation yields slight gains (without ROVER)
- **Arabic-English examples**

wrfD Hmzp k\$F Asm Al\$Rkp Al<srA}ylyp wtHdyd mA <*A kAn TlbhA qd rfD >w qbl. Hamza refused to disclose the name of the Israeli company, and to determine whether the request had been rejected or before.	Ref: Hamza refused to disclose the name of the Israeli company and to reveal whether its application was accepted or refused.
wH*r AlwzrA' AlmjtmEwn bTlb mn flsTyn mn AstmrAr Al>n\$Tp AlnwWyp Al<srA}ylyp... the ministers warned the assembled at the request of the Israeli nuclear activities outside ...	Ref: The meeting ministers warned, following a request by Palestine, against the continuation of the Israeli nuclear activities outside...

- **Main challenges**
  - Reordering. Allowing any phrase movement is not feasible, and constraints produce wrong word order
  - Fluency. Translation often omits connectors
  - Need for more structure (ie. syntax) ?