Automatic Speech Recognition of Agglutinative Language based on Lexicon Optimization

米吉提•阿不里米提

Mijit Ablimit

2014.11.03，清华大学 FIT 楼 1-303

Outline

1. Introduction
2. Review of lexicon optimization methods
3. Morphological segmenters for Uyghur language;
   ASR results based on various morphological units
4. Morpheme concatenation approaches by comparing two layers of ASR results
5. Discriminative lexicon optimization approaches
6. Comparison of lexicon optimization methods

Chapter 1:

Introduction

Prediction

- We predict everything in our life.

  - Tomorrow
  - Weather
  - Future
  - Speech
  - People
  - ……..

- Not always is accurate. Bu we can improve accuracy.

$$P(Y \mid X)$$

Automatic Speech Recognition (ASR):

- transcribing speech into text

$$\hat{S} = \underset{S}{\operatorname{argmax}} P(S|U)$$

- U is the acoustic features of speech
- S is the corresponding text

Bayes's law is used to decompose into:

$$\hat{S} = \underset{s}{\operatorname{argmax}} \frac{P(U|S)P(S)}{P(U)} = \underset{s}{\operatorname{argmax}} P(U|S)P(S)$$

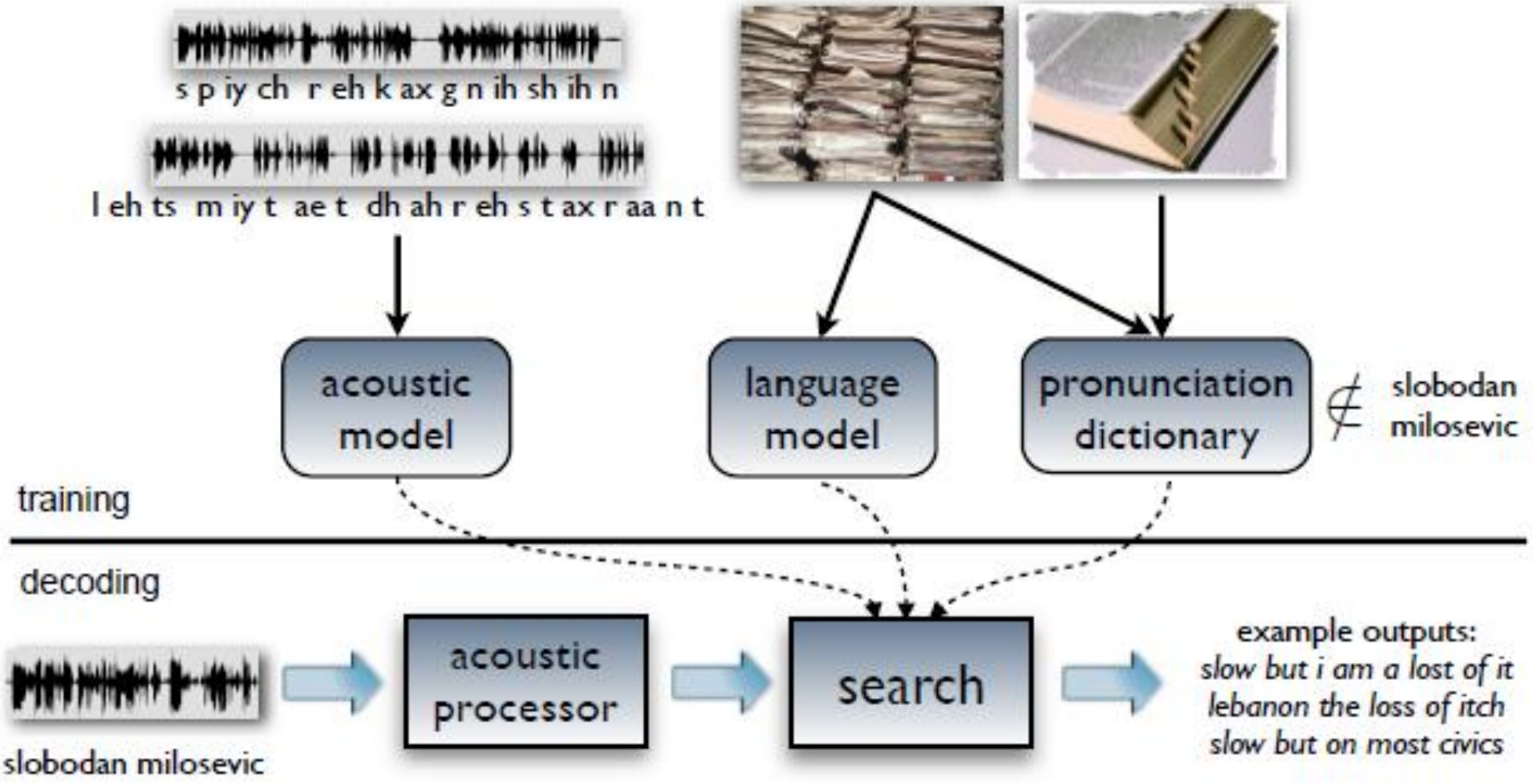
- P(U|S) difficult to train, because of the data sparseness.

- Instead, acoustic model (AM) is trained to recognize the smallest linguistic sequence: phonemes X

The practical formulation for ASR:

$$\hat{S} \approx \underset{S,X}{\operatorname{argmax}} P(U|X)P(X|S)P(S)$$

- P(U|X) Acoustic Model (AM)
- P(X|S) Lexical Model
- P(S) Language Model (LM)


s p iy ch r eh k ax g n ih sh ih n
l eh ts m iy t ae t dh ah r eh s t ax r aa n t
acoustic model
language model
pronunciation dictionary
slobodan milosevic
training
decoding
acoustic processor
search
slobodan milosevic
example outputs:
slow but i am a lost of it
lebanon the loss of itch
slow but on most civics

LM P(S) providing linguistic constraints

$$P(S) = P(t_1 t_2 \dots t_N) = \prod_{i=1}^{N} P(t_i | t_1 \dots t_{i-1})$$

- Test (or sentence) S can be expressed by smaller units.
- Product rule conditioned upon previous unit sequence.
- Still the dimensionality is infinite.

n-gram language modeling

$$P(t_1 t_2 \dots t_N) \approx \prod_{i=1}^{N} P(t_i | t_{i-n+1}^{i-1})$$

- Each unit is only dependent on the previous (n-1) units.
- Practically, up until 3-gram or 4-gram is adopted.
- Choosing the lexical unit is very importance first step.

Possible lexical units for LM

- Word
- Morpheme
- Syllable
- Statistical morphemes （pseudo-morpheme） （quasi-morpheme）

Evaluations for ASR system

- **Word Error Rate (WER)**, Morpheme Error Rate (MER), CER

$$WER = \frac{insertions + deletions + substitutions}{words\ in\ reference}$$

- **Lexicon size**,

an optimal lexicon set should maintain:

high coverage while maintaining small size

Evaluations for ASR system, Perplexity

- Perplexity, is calculated on the basis of modeled units
- For $S = t_1 t_2 \ldots t_N$ $Perplexity = 2^{H(S)}$
- For fare comparison on different units

$$Normalized\ PP^* = PP^{\frac{N_t}{N_w}}$$

- Where Entropy : $H(S) = -P(S)\log_2 P(S)$
- Cross Entropy: $H(S) = -P(S)\log_2 P(S|model)$

Problems: Limits in Modeling

- Formulation is on infinite training data,

$$P(t_1 t_2 \dots t_N) \approx \prod_{i=1}^{N} P(t_i | t_{i-n+1}^{i-1})$$

- Limited training data reduces model reliability.
- Increasing size of **n** may decrease the model reliability.
- the optimized lexicon set is simply equivalent to a flexible n-gram model

$$Length\left(W_{i-n+1}^{i-1}\right) > Length\left(M_{i-n+1}^{i-1}\right)$$

Problems: Linguistic constraints

- Speech is different from person to person, from time to time.
- These changes include

phonetic changes : **phonetic harmony or disharmony.**
morphological changes: **omission, insertion, substitute.**

- Called **co-articulation** effect.
- Difficult to extract manually. And manual word *may not* fit for actual speech.

Goal of this research:

Optimal unit set which

- Considering both

  statistical parameters and **linguistic constraints**.

- Reducing both

  Lexicon size and **WER**

- Directly linked to the ASR accuracy, automatically reduce **co-articulation** problem.

Part 2:

Review of Lexicon optimization methods

Data driven approaches

- Merging short and frequently co-occurred units.
- Mutual information (MI) can provide a threshold.

$$LM(m_i\ m_j) = \sqrt{\mathrm{P_f}(m_i|m_j)\mathrm{P_r}(m_j|m_i)} = \frac{\mathrm{P}(m_i, m_j)}{\sqrt{\mathrm{P}(m_i)\mathrm{P}(m_j)}}$$

Statistical modeling for concatenation

- Recognition correctness is formulated on unit frequency, length .
- Perplexity is calculated for basic phoneme units as a criterion for building new units.
- Sub-word units is used to reduce and detect OOV , and online learning of OOV is adopted.

Unsupervised lexicon extraction

- Discovery lexicon from untagged corpora.
- Maximum Description Length (MDL) is utilized for different structures.
- Bayesian framework utilized for unit selection to overcome overfitting problem of Maximum Likelihood estimation.

Unsupervised lexicon extraction

- Maximum a posteriori (MAP) formulation used for sub-word segmentation.
- Frequency and Length properties are utilized.

$$argmax\, P(M|corpus) = argmax\, P(corpus|M)P(M)$$

$$P(M) = M!\ P(\,freq(t_1)\ldots freq(t_N)\,)\cdot\prod_{i=1}^{N}[(1-P(\#))^{length(t_i)}\cdot P(\#)\cdot\prod_{j=1}^{length(t_i)}P(c_j^{t_i})]$$

Conclusions

- Summery
  - Data driven approaches
  - Statistical modeling for unit concatenation
  - Unsupervised segmentation

- Actually based on occurrence Frequency and Length.
- Not considering linguistic constrains.
- Not strong relation with ASR performance.

Part 3:

Sub-word segmentation in Uyghur language and Baseline ASR results

Outline

- **Morpheme segmentation**
- **Language models on different units**
- **Experimental results of ASR**

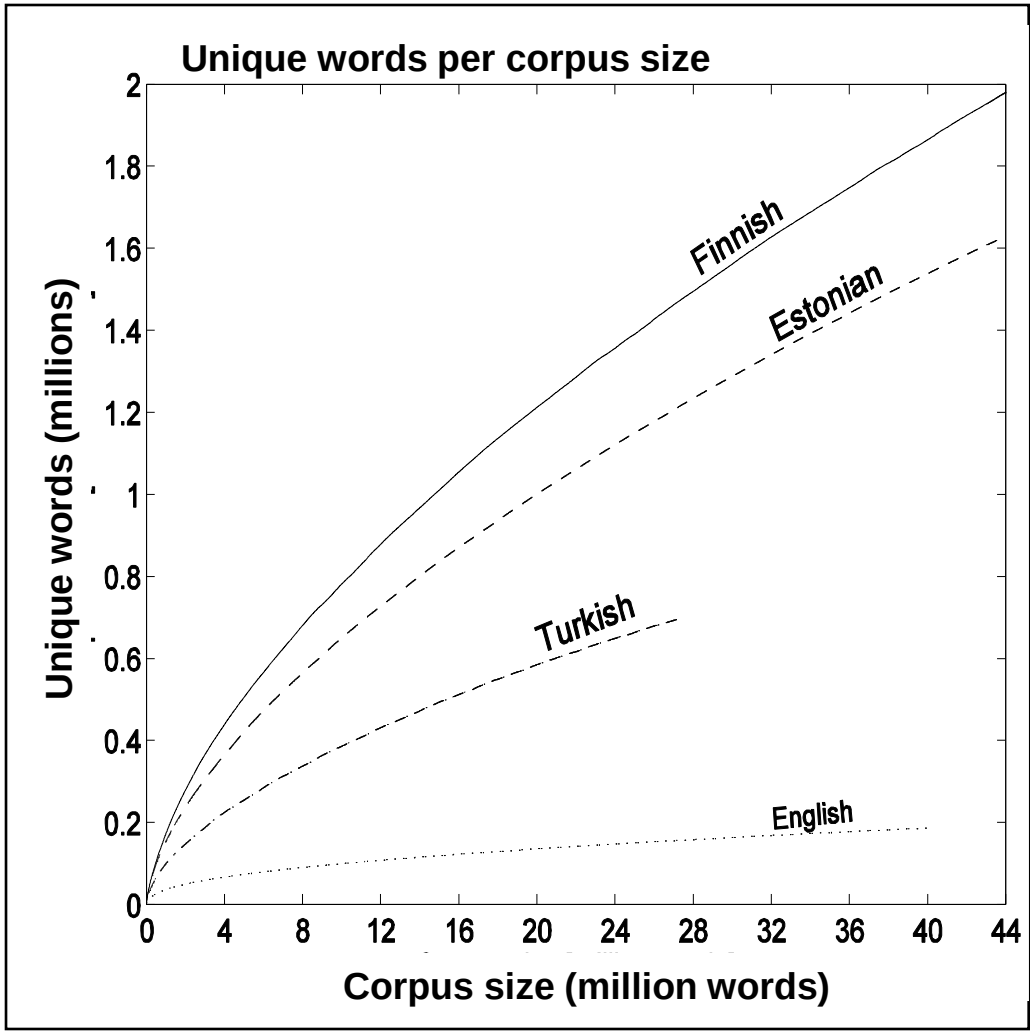
Uyghur language

- **Uyghur language is an agglutinative language, belongs to Turkish Language Family of Altaic Language system.**
- **Agglutinative and highly-inflected languages suffer from a severe vocabulary explosion.**
- **Sentences in Uyghur consist of words, which are separated by space or punctuation marks.**
- **Smaller units are considered to be a good option in many inflectional languages like Arabic, Turkish, Persian, Finish, German, Korean…**

The vocabulary problem

- ***Agglutinative* and highly-inflected languages** suffer from a severe **vocabulary explosion**

- **Theoretically *Vocabulary Size* in Uyghur is infinite**



From: Unsupervised Morpheme Analysis – *Overview of Morpho Challenge 2007 in CLEF*

Uyghur language and morphology

- *Uyghur language is an agglutinative language, belongs to Turkish Language Family of Altaic Language system.*

Müshükning kəlginini korgən chashqan hoduqup qachti.
(ねこが きたのを みた ねずみ(が) おどろいて にげた)
(The mouse who saw the cat coming *was startled and escaped.*)
words are separated naturally

- *morpheme sequence: format* ***" prefix + stem + suffix1 + suffix2 + … "***

Müshük+ning kəl+gən+i+ni kor+gən chashqan hoduq+up qach+ti.
(ねこ-が き-た-の-を み-た ねずみ-(が) おどろい-て にげ-た)
Suffer from phonological morphological changes

- *syllable sequence: format* ***"CV[CC]"*** *(C: consonant; V: vowel)*

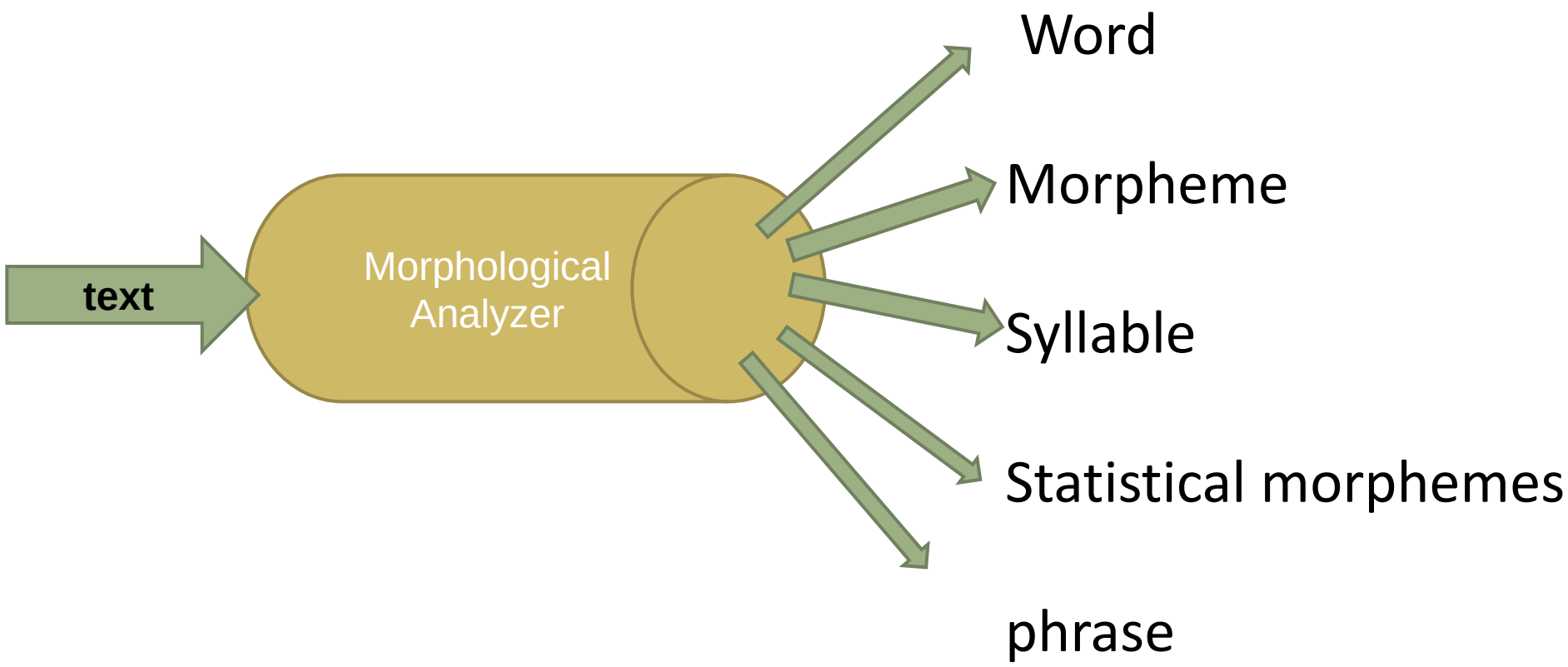
Mü+shük+ning kəl+gi+ni+ni kor+gən chash+qan ho+du+qup qach+ti.

Problems during morpheme segmentation

- **insertion, deletion, phonetic harmony, and disharmony (vowel assimilation, vowel weakening).**

- 1) assimilation should be recovered to standard surface forms.
  almiliring=alma+lar+ing
- 2) morphological change, which is deletion and insertion.
  oghli= oghul + i ; binaying=bina+[y]+ing
- 3) phonetic harmony.
  Kyotodin= Kyoto + din; Newyorktin= Newyork + tin
- 4) ambiguity.
  berish= bar(go/have)+ish, berish= bər(give)+ish

Morphological Analyzer



Supervised morpheme segmentation

- *Statistical modeling*

- A statistical model can be trained in a fully supervised way. A text and its manual segmentation is prepared.

  A text corpus of 10025 sentences, collected from general topics, and their manual segmentations are prepared.

	tokens	vocabulary
word	**139.0k**	**35.37k**
morpheme	**261.7k**	**11.8k**
character	**936.8k**	
sentence	**10025**	

- More than 30K stems are prepared independently and used for the segmentation task.

Probabilistic model for morpheme segmentation

- Intra-word bi-gram probabilistic formulation is:

$$\begin{cases} \mathbf{P(stem,\ firstSuffix)} \\ \mathbf{P'(stem)P(anySuffix \mid stem)} \quad \text{for smoothing} \end{cases}$$

in which

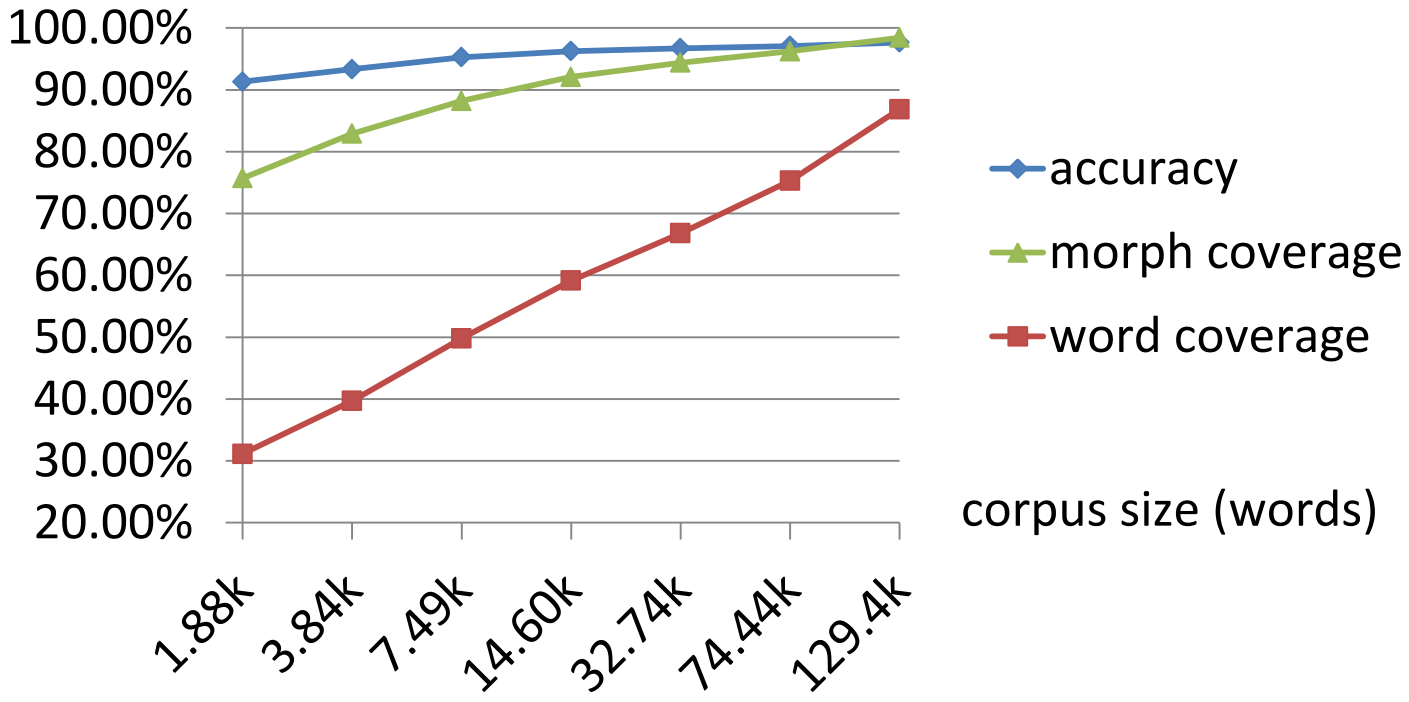
$$\mathbf{P'(stem)} = \frac{\mathbf{stemFrequency}}{\mathbf{(stemToken+stemVocabulary)}}$$

$\mathbf{P(anySuffix \mid stem)}$ probability of a stem linked with a suffix

Surface realization is considered. Standard morpheme format is exported.

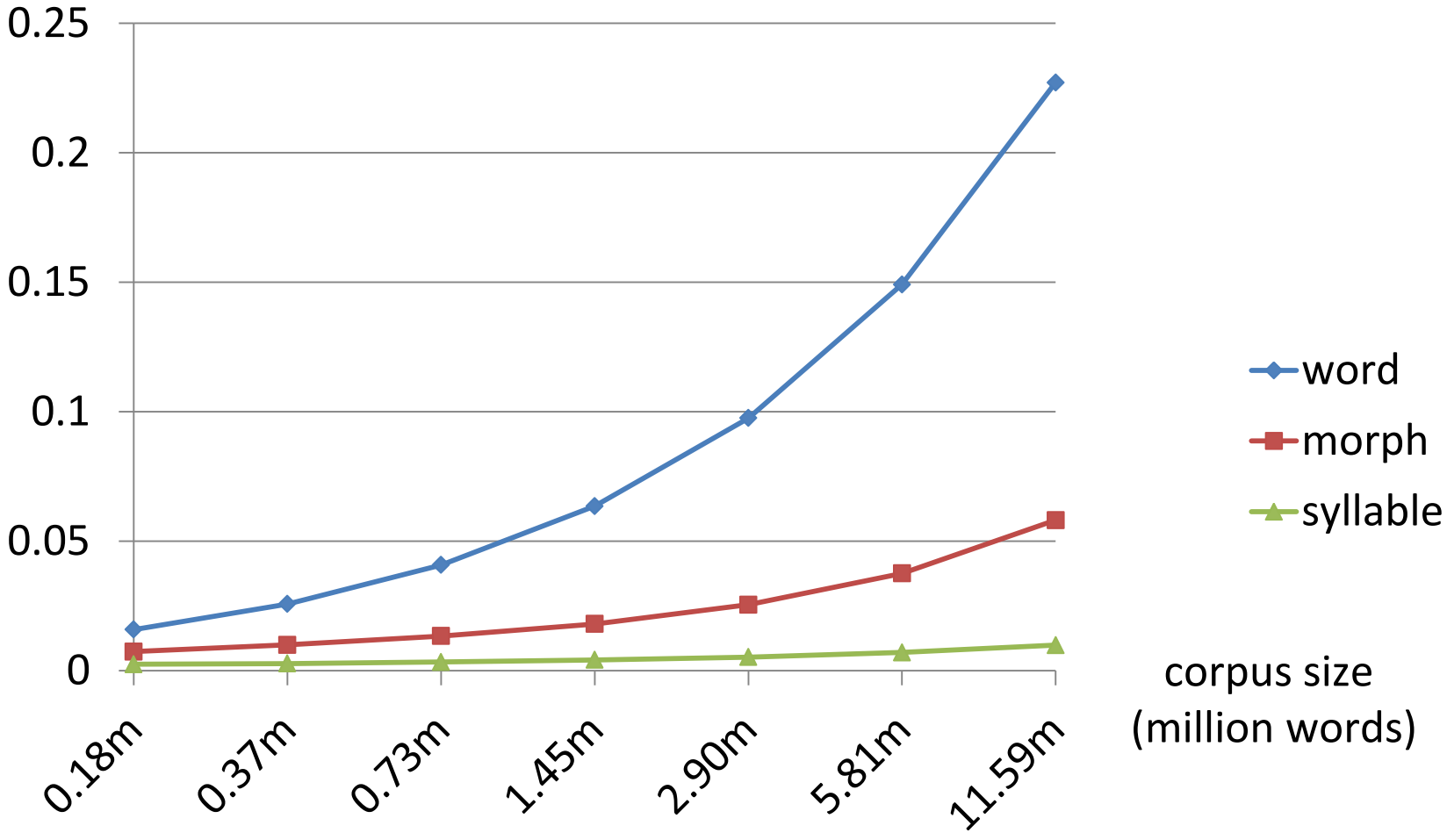
- For a candidate word, all the possible segmentation results are extracted before their probabilities are computed to get the best result.

Morpheme segmentation accuracy and coverage



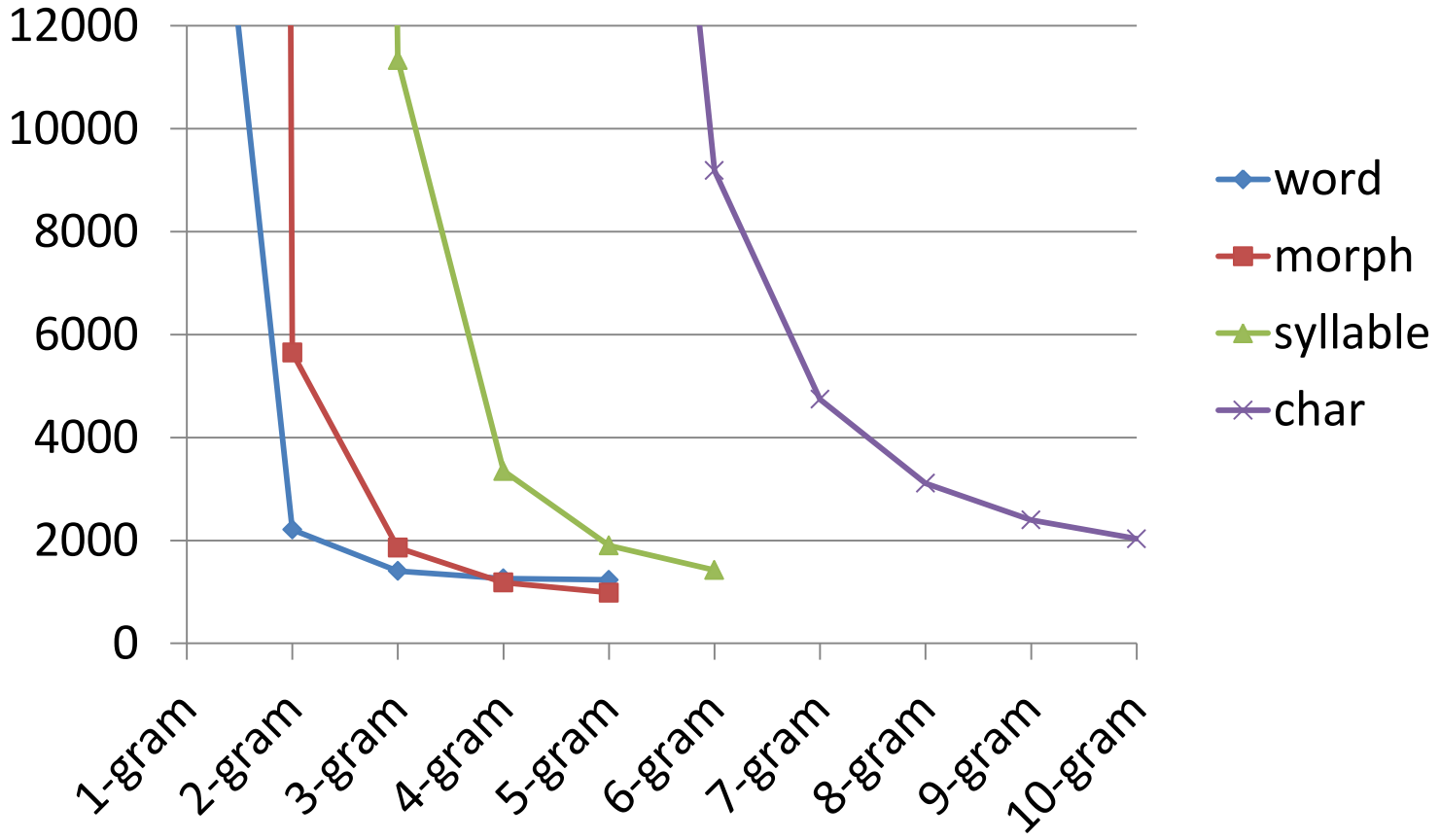
- Word coverage is 86.85%. Morpheme coverage is 98.44%.
- The morpheme segmentation accuracy is **97.66%**

Vocabulary comparison of various units



Vocabulary size explode by words

Perplexity comparison of various n-grams, normalized by words



- Perplexities by various unit sets will converge to similar results.
- Slight gain by longer units with smaller size of n.
- Morpheme slightly outperformed word, because of small OOV rate.

Uyghur ASR experiments

-*Uyghur Acoustic Model*

- Training speech corpus is selected from general topics. And used for Uyghur ***acoustic model (AM)*** building.

- Test corpus is independent from the training speech corpus

Corpus	Unique sentences	Speakers	Female	Male	Age	Total utterance	time (hour)
Training	13.7K	353	187	166	19-28	62k	158.6
Test	550	23	13	10	22-28	1468	2.4

Uyghur ASR experiments

-*Uyghur Acoustic Model*

- The text corpus of 630K sentences for language modeling.
- The sentences are segmented to word, morpheme, and pseudo-morpheme units, and LMs are separately constructed based on each of them.
- An acoustic model based on tri-phone HMMs with 3000 shared states and 16 Gaussian mixtures was trained for 34 Uyghur phones (8 vowels, 24 consonants, and 2 silence models). The acoustic features consist of 12 MFCCs, ΔMFCCs and ΔΔMFCCs together with Δpower and ΔΔpower.

ASR systems based on various morphological units

Four different language models are built.

1) Word based model
2) Morpheme based model
3) Stem-Suffix (word endings) based model
4) Syllable based model

LM names	Word	Stem-Suffix	morph-3gram	morph-4gram	morph-5gram
Vocabulary	227.9k	74.5k	55.2k	55.2k	55.2k
Morph Error Rate(%)	18.88	21.69	22.73	21.64	22.98
Word Error Rate (%)	25.72	28.13	28.96	27.92	29.31

The syllable vocabulary is 6.58k and the *syllable error rate* is 28.73%.

Word-based ASR result is automatically segmented to morphemes and syllables. Corresponding MER is 18.88%, SER is 15.42%.

Conclusion

- Supervised morphological unit segmentation achieved 97.6% for Uyghur language.
- Morpheme provides syntactic and semantic information which is convenient for ASR and NLP researches.
- Uyghur LVCSR system on various linguistic units are build for the first time.
- Longer units (word) outperform other sub-word units in ASR application.

Chapter 4:

Morpheme Concatenation Approach based on feature extraction from two layers of ASR results

Outline

- Corpora and Baseline systems
- Problematic sample extraction
- Experimental results

Aligned ASR results of word and morpheme units

reference word	Yash		cheghinglarda		bilim		elishinglar		kerək	
reference morph	Yash		chegh_ing_lar_da		bilim		el_ish_ing_lar		kerək	
word ASR result	Yash O		cheghinglarda O		bilim O		berishinglar X		kerək O	
morph ASR result	Yash O		chegh_ing_da X		bilim O		el_ish_ing_lar O		kerək O	

- Word unit provides better ASR performance, vocabulary size explode, and causing OOV.
- Morpheme unit smaller vocabulary size, high coverage, but often short and easily confused.

ASR results on various morphological units

Baseline models		WER (%)	Vocabulary size
Morph. 4-gram	cutoff-2	27.92	55.2k
	cutoff-5	28.11	27.4k
Word 3-gram	cutoff-2	25.77	227.9k
	cutoff-5	26.64	108.1k
FMS-500		28.14	274.9k
Stem & word endings		28.13	74.5k

Best ASR results are aligned for sample extraction

Frequent Morpheme Sequence (FMS), Co-occurred less than 500 times are merged.

- Morpheme 4-gram generate best results for morpheme based LMs.
- Cutoff-F means units whose frequency is less than F are considered *UNKNOWN* .

Comparing ASR results of word and morpheme units

- We extract useful patterns from the ASR results. Analyze reasons for the confusion.

Main reasons for misrecognition	Examples (English translation)
Phonetic harmony or co-articulation	yigirmə-yigirmi (twenty), vottura-votturi (middle)
Confusion in frequent short stems with many derivatives	biz, vu, bash, yər (we, he/her, head, land)
Phonetic similarity	həmmə-əmma (all, but)
Ambiguity	uni-u+ni (he, him)
Too many suffix insertions	ish+lap+p+i+ish+ni

The co-articulation problem can be partly solved by merging units.

- We focus on reasons for confusion, and ascribed to several features: *error frequency, length, and attribute (stem or word-ending).*

Features for extracting samples from ASR results

- The patterns we considered are three types of features.

  - First , the error frequency of word unit.

$$\Phi_{freq}(w) = \begin{cases} \text{true} & if \;\; w \;\; misrecognized \; more \; than \; twice \\ \text{false} & otherwise \end{cases}$$

  - Second, the length of morphemes.

$$\Phi_{length}(m_i) = \begin{cases} \text{true} & if \;\; length(m_i) \;\; is \; less \; than \; 2 \\ \text{false} & otherwise \end{cases}$$

  - Third, Attributes (stem and word-ending)

$$\Phi_{stem}(st_i) = \begin{cases} \text{true} & if \; stem \;\; st_i \;\; misrecognized \; more \; than \; 10 \; times, \\ & \quad and \; length \; is \; less \; than \; 4 \; syllables \\ \text{false} & otherwise \end{cases}$$

$$\Phi_{word\_ending}(we_i) = \begin{cases} \text{true} & if \; word\_ending \;\; we_i \;\; misrecognized \\ & \quad when \; connected \; with \; a \; short \; stem \;\; st_i \\ \text{false} & otherwise \end{cases}$$

stem and word-ending features are special for *Agglutinative* languages.

Experimental evaluation for

- *error frequency feature*

- Word candidates are selected according to these features. And added to lexicon of morpheme unit.

➢ Iteratively application of error frequency feature shows accumulative improvements.

Iterations	Baseline	First round	Second round
WER(%) on training data	31.95	28.62	27.01
WER(%) on test data	28.11	26.11	25.82
Vocabulary size	27.0k	40.3k	46.0k

- When we extract misrecognized words from the test set, we found that only 50% of them are covered by the training data set.

Experimental evaluation for

- *different features*

- The effects of various features and their combined effects.

Models	WER(%)	△WER (%)	Vocabulary size
Morpheme-based baseline	28.11	-	27.3k
Error frequency feature	26.11	2.00	40.3k
Length feature	27.19	0.92	32.8k
Attribute features	26.74	1.36	36.3k
Attribute + Length features	25.80	2.31	41.2k
Attribute + Length + Error freq features	24.89	3.22	56.7k

The features have accumulative effects.

- The result is significantly outperformed both of the baseline models.

Conclusion

- We have proposed a manual feature extraction approach based on two layers ASR result comparison.
- Instead of speculating linguistic or statistical properties, we directly analyze the ASR results and identify useful features.
- The proposed method significantly reduced both WER and lexicon size compared to the best word-based model.
- Directly related with ASR results, no direct link with OOV.

Chapter 5:

Discriminative Lexicon Optimization Approach for ASR

Outline

- Feature extraction from the aligned ASR results
- Discriminative approaches for lexicon optimization
- Lexicon design and baseline ASR systems
- Experimental evaluations

Aligned ASR results of morpheme & word based models

Means: Study hard when you are young.

reference word	Yash	cheghinglarda	bilim	elishinglar	kerək
reference morph	Yash	chegh_ing_lar_da	bilim	el_ish_ing_lar	kerək
word ASR result	Yash O	cheghinglarda O	bilim O	elishinglar O	kerək O
morph ASR result	Yash O	chegh_ing_da X	bilim O	el_ish_ing_lar O	kerək O

CRITICAL CASE

We automate the manual feature extraction approach

Sample extraction from the aligned ASR results

- The CIRITICAL samples are extracted.

Word	≠	Morph.	percentage
X		X	68%
O		X	28.5%
X		O	3.5%

⇨ Naïve method =error frequency sampling

- Naïve method:
  - Misrecognized morphemes are merged into words.
  - WER greatly reduced with this experiment.
  - Difficult to cover all words.

Feature extraction from the aligned ASR results

- Given all the training sample pairs: $(w = m_1 m_2 \ldots)$
- We can extract binary features: $\Phi(w^i)$ — *i*: sample index
- And desired value:

$$y^i = \begin{cases} +1 & \text{if CRITICAL\_CASE is true} \\ 0 \text{ or } -1 & \text{otherwise} \end{cases}$$

w	$m_i\, m_j \ldots$	word ≠ morph.
cheghinglarda	chegh_ing_lar_da	O X
$\Phi_{\text{unigram\_}lar}(cheghinglarda) = 1$		$y^i = 1$

$$\Phi_{\text{bigram\_}m_i\, m_j}(w) = \begin{cases} 1 & \text{if } morph.bigram\,(m_i\, m_j)\ exists\ in\ w \\ 0 & \text{otherwise} \end{cases}$$

Feature extraction from the aligned ASR results

- ***unsupervised*** utilizes all CRITICAL_CASE as *Word ≠ Morph.*

Word ≠ Morph.		percentage
X	X	68%
O	X	28.5%
X	O	3.5%

unsupervised

supervised

- All the training binary samples are extracted: $(\Phi(w^i),\ y^i)$

$$(i = 1, \ldots l, \Phi(w^i) \epsilon \{0,1\},\ y^i \epsilon \{0, +1\})$$

- These samples are feed to machine learning algorithms:
  - Perceptron , SVM, LR

Discriminative approach - *Perceptron*

- For the ***perceptron***, we define an evaluation function:

$$f(w^i) = \sum_s \Phi_s(w^i)\alpha_s = \Phi(w^i)\,\alpha$$

α_s is a weight for the feature $\Phi_s(w^i)$

- The standard sigmoid function is applied to the linear estimation function.

$$g(w) = \frac{1}{1+e^{-f(w)}} \qquad g'(w)|_{f(w)} = g(w)(1-g(w))$$

- The weight vector is updated as:

$$\alpha = \alpha + \eta\, g'(w^i)(y^i - g(w^i))\Phi(w^i)$$

Easily converging into a local optimum with a large dimension of features

Discriminative approach – *SVM & LR*

- Both methods solve the following unconstrained optimization problem

$$min_{\alpha} \frac{1}{2} \alpha^T \alpha + C \sum_{i=1}^{l} \xi(\alpha; \Phi(w^i), y^i)$$

with different loss functions:

$$\xi(\alpha; \Phi(w^i), y^i)$$

- For SVM, the loss function is:

$$\xi(\alpha; \Phi(w^i), y^i) = \max(1 - y^i \alpha^T \Phi(w^i), 0)^2$$

- For Logistic Regression , the loss function is:

$$\xi(\alpha; \Phi(w^i), y^i) = \log(1 + e^{-y^i \alpha^T \Phi(w^i)})$$

Lexicon Design

- These features are then generalized to all units in the text corpus.
- The ***concatenation*** process is repeated in sequence while the condition *(g(w)>0.5)* is met.
- Can be applied to sub-word within word boundary; search is done while the ***condition*** is met.
- 4-gram LM and Cutoff-5 are used for all experiments.

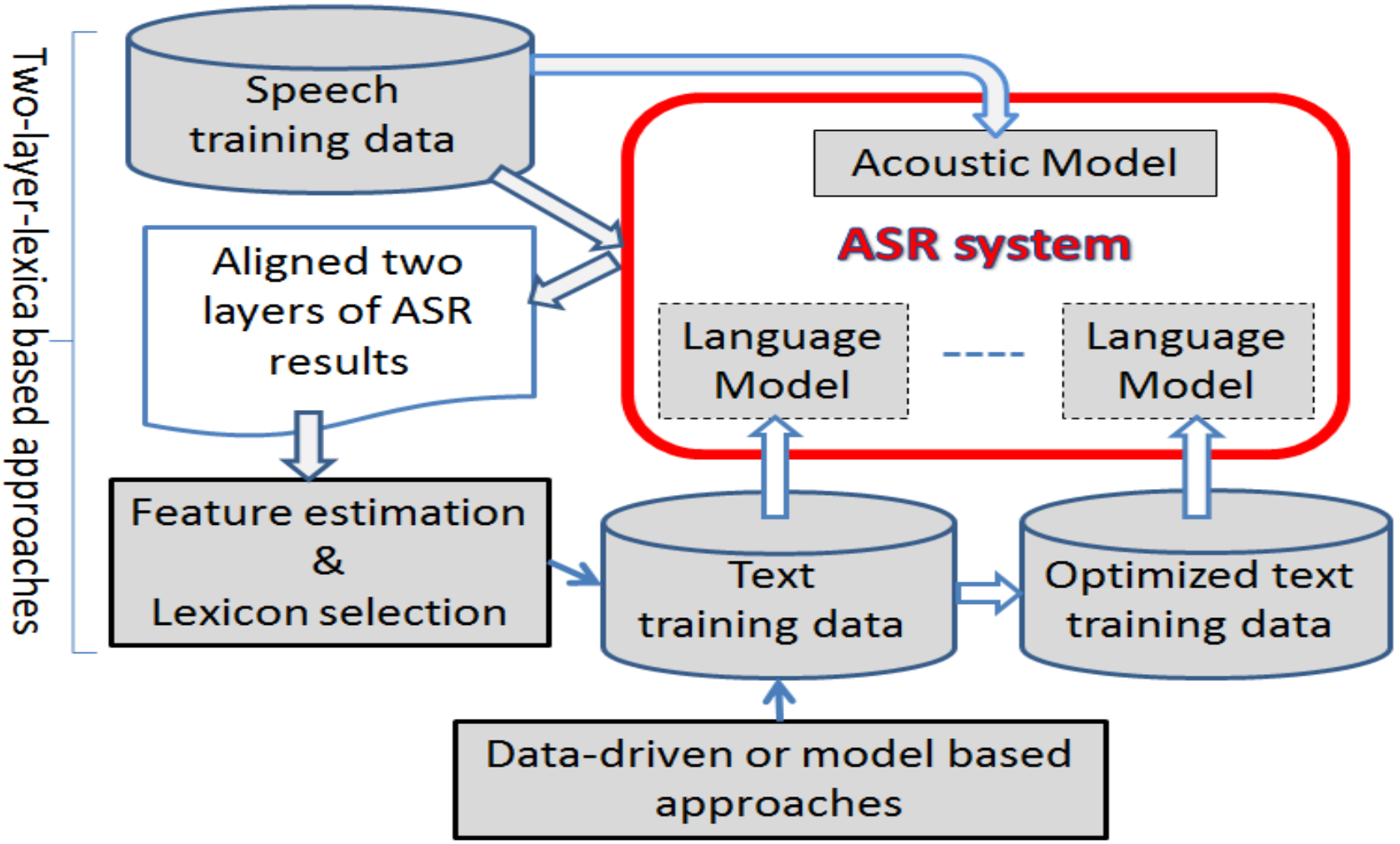
Baseline ASR results with various units

Baseline models		WER (%)	Vocabulary size	OOV
Morph 4-gram	cutoff-2	27.92	55.2k	0.3%
	cutoff-5	28.11	27.4k	0.7%
Word 3-gram	cutoff-2	25.72	227.9k	2.8%
	cutoff-5	26.64	108.1k	4.4%

Best ASR results are compared for feature extraction

- Outlier samples are removed when the frequency of the CRITICAL samples are less than a ***filtering threshold N***
- The *N-gram feature dimension* covered by the ***speech training corpus*** are :
  - unigram features: 17K
  - bigram features : 53K

Flow chart of discriminative approach



Effect of sample filtering threshold with unigram feature

threshold		N=0	N=1	N=2	N=3	N=4	N=5
perceptron	WER (%)	26.69	25.93	**25.87**	26.18	26.28	26.54
	Lexicon size	104.5K	90.2K	74.8K	63.6K	55.3K	50.1K
LR	WER (%)	25.99	**25.57**	25.91	25.93	26.01	26.22
	Lexicon size	102.4K	91.2K	79.9K	70.1K	62.4K	56.5K
SVM	WER (%)	**26.05**	26.03	25.93	25.93	26.00	26.22
	Lexicon size	103.4K	94.6K	83.7K	73.5K	65.4K	59.2K

- **SVM and LR are more robust against less reliable samples.**

Comparison of results on different units and features

Units		Word		Sub-word	
Features		unigram	bigram	unigram	bigram
perceptron	WER (%)	25.87	25.99	25.96	25.27
	Lexicon size	74.8K	67.3K	40.7K	49.9K
LR	WER (%)	25.99	25.75	25.77	24.87
	Lexicon size	102.4K	85.4K	44.0K	65.8K
SVM	WER (%)	26.05	25.86	27.05	**24.61**
	Lexicon size	103.4K	80.1K	34.7K	**55.1K**

- ***Sub-word*** optimization significantly reduce both WER and lexicon size.

Supervised and Unsupervised feature extraction

		supervised	unsupervised
perceptron	WER (%)	25.55	25.27
	lexicon size	49.7K	49.9K
LR	WER (%)	25.34	24.87
	lexicon size	46.3K	65.8K
SVM	WER (%)	25.42	24.61
	lexicon size	45.1K	55.1K

Word ≠ morph.		percentage
X	X	68%
O	X	28.5%
X	O	3.5%

sub-word bigram feature

- The ***unsupervised*** training is scalable to a large speech data.

Summary of the results

Models		WER(%)	Lexicon size	OOV
baseline morpheme		28.11	27.4k	0.7%
baseline word		25.72	227.9k	2.8%
best MI method		25.60	53.3k	0.7%
SVM sub-word bigram feature	cutoff-2	24.64	101.2k	0.7%
	cutoff-5	24.61	55.1k	0.9%

- The optimized system is very ***stable.*** Not much effected by Cutoff rate.
- ***SVM & LR*** are more robust than ***perceptron***.

Conclusion

- A **novel** discriminative approach to lexicon optimization for highly inflectional languages.
- Automatically optimize lexical units, outperformed other methods with smallest WER and lexicon size.
- ***SVM & LR*** are more effective than ***perceptron***.
- Sub-word optimization based on bigram feature produce the best result .
- Can be trained on un-transcribed speech data.

Chapter 6:

Comparison of lexicon optimization methods

Introduction

- Both ***supervised*** and ***unsupervised*** **segmentation** methods, and ***manual*** and ***automatic*** **concatenation** methods are compared for ASR.

- The ***unsupervised segmentation*** methods can split words into **morpheme-like units** from a raw text corpus.

Various approaches	manual	automatic
Segmentation	Supervised linguistic morpheme based approach	Unsupervised morphs based approach
Concatenation	manual extraction of problematic morpheme sequences	Discriminative approach
		Statistical concatenation approach

Unsupervised lexicon extraction

- Maximum a posteriori (MAP) formulation used for sub-word segmentation.
- Frequency and Length properties are utilized.

$$argmax\, P(M|corpus) = argmax\, P(corpus|M)P(M)$$

$$P(M) = M!\ P(\,freq(t_1)\ldots freq(t_N)\,)\cdot\prod_{i=1}^{N}[(1-P(\#))^{length(t_i)}\cdot P(\#)\cdot\prod_{j=1}^{length(t_i)}P(c_j^{t_i})]$$

Segmentation based ASR results

Segmentation based LM		WER (%)	vocabulary size	OOV
morpheme 4-gram	cutoff-2	27.92	55.2k	0.3%
	cutoff-5	28.11	27.4k	0.7%
word 3-gram	cutoff-2	25.72	227.9k	2.8%
	cutoff-5	26.64	108.1k	4.4%
Statistical morph 4-gram	cutoff-5	25.01%	94.5k	0.9%
	cutoff-2	25.04%	131.3k	0.8%

MAP model is used for unsupervised segmentation.
(Chapter 2, slide 14)

- Statistical morphs are not considering linguistic information, have a statistical properties comparible to word units

and

- Have a competitive result to the discriminative method, but with a larger lexicon size.

Statistical model based concatenation approach

4-gram models		WER(%)	vocabulary size	OOV
Linguistic morpheme	Cutoff-5	28.11	27.4k	0.7%
	Cutoff-2	27.92	55.2k	0.3%
Statistical optimization	Cutoff-5	24.96	98.35k	0.9%
	Cutoff-2	24.85	139.0k	0.8%

Morfessor tool (Creutz) is modified to concatenate *linguistic morphemes* into sub-words.

- The concatenative approach is based on linguistic morphemes.
- Directly concatenated from a morpheme based text corpus, with the statistical approach based on MAP. (Chapter 2, slide 14)

Comparison of Segmentation and concatenation approaches

approaches		WER (%)	Lexicon size
Word-based baseline best result (3-gram)		25.72	227.9k
segmentation	Supervised linguistic morpheme based approach	27.92	55.2k
	Unsupervised morphs based approach	25.01	94.5k
concatenation	Statistical concatenation approach	24.96	98.35k
	Manual extraction of problematic morpheme sequences	24.89	56.7k
	Discriminative approach, SVM based, automatic approach	24.61	55.1k

Automatic and manual concatenation methods are compared.

Cutoff-5 and 4-gram LMs

- Linguistic morphemes are used for all concatenation approaches, significantly decreased WER.

Conclusion

- A direct concatenation method based on statistical model proved to be very effective for reducing WER.
- All *supervised* and *unsupervised* segmentation approaches are compared for ASR.
- Several different concatenation approaches for morphemes are compared.

Concluding remarks

- For the first time, Uyghur language based morphological analyzers and ASR systems for various morphological units are investigated and optimized.
- Supervised and unsupervised morpheme segmentation methods and ASR applications are compared.
- A novel approach based on comparing two layers of ASR results are proposed.
- The proposed discriminative approach has significantly reduced both WER and lexicon size.
- The optimized lexicon based ASR system is very stable compared to baseline systems.

Discussion on generality of the proposed approach

- The basic assumption is word unit outperform *predefined* morpheme unit. There are similar tendencies in other languages like Turkish, Korean.

- The predefined morphemes, for some languages, may have a good ASR performance, thus does not satisfy the basic assumption.

  - Units in Japanese are not clear, the extracted lexicon already be the optimal.
  - Rule based or statistical approaches already optimized the lexicon.

- Larger training data size of resource rich languages benefits the longer units, which is consistent with our basic assumption.

Discussion on generality of the proposed approach

- DNN based AM also proved the effectiveness of the optimized lexicon ,
- However, overthrown the basic assumption of the two-layer-lexica based approach.

LMs for DNN AM	WER(%)	Lexicon size
word based	16.50	227.9K
morpheme based	14.50	27.4K
optimized pseudo-morph.	12.89	55.1k

Thank you