Trigger-Based Language Model Adaptation for Automatic Transcription of Panel Discussions

Carlos Troncoso Alarcón

Speech and Acoustics Laboratory
NAIST

April 27, 2006

Background

- Conventional *n*-gram LM [Bahl '83]
  - Powerful for modeling short-distance dependencies
  - Unable to model dependencies longer than *n* (*n* = 2~4)

  e.g. I used to go to this resort on the beach with…

  P_{3GRAM}
- Alternative LMs
  - Short distance: • Class *n*-gram [Brown '92] • Mixture-based LMs [Iyer '99]
  - Intermediate distance: • Long distance *n*-gram [Huang '93]
  - Long distance: • Cache-based LM [Kuhn '92] • Trigger-based LM [Rosenfeld '96] • LSA-based LM [Bellegarda '00]

Trigger-Based LM

- Trigger pairs
  - Semantically correlated word pairs (resort → beach)
  - *A* → *B* means *"A* 'triggers' the appearance of *B"*
  - Constructed from large corpus using average mutual information (AMI) within a text window
- Raise probability of words triggered by others

★ Able to model dependencies longer than *n*

$P_{TRIGGER}$

e.g. I used to go to this resort on the beach with…

P_{3GRAM}

Limitations of Conventional Trigger-Based LM

- Constructed from text window
  - Window limits scope of dependencies the model can capture ⇒ Local constraints

  ⇒ Global topic constraints by TF/IDF
- Most potential lies in "self-triggers"

  (e.g. beach → beach)
  - Self-triggers virtually equivalent to cache-based LM ⇒ Small improvement

  ⇒ Effective use of non-self-triggers
- So far applied to written language (newspapers)
  - Corpora too general in topic ⇒ Task dependency lost

  ⇒ Trigger-based LM adaptation to target domain

Application to Conversational Speech

- Conversations and meetings usually centered in a topic

  ⇒ Trigger pairs capture long-distance topic constraints
- Problems of conversational speech
  - Disfluencies (filled pauses, repetitions, repairs…)
    - Sentences can become ungrammatical
    - Disfluencies contribute to data sparseness
    - Longer dependencies between words

    ⇒ Trigger-based LM insensitive to disfluencies
  - Small amount of available in-domain data
    - Conversational text corpora expensive to produce
    - Insufficient to derive reliable task-dependent models
    - Web-based approaches not domain matched

    ⇒ Effective training of trigger-based LM

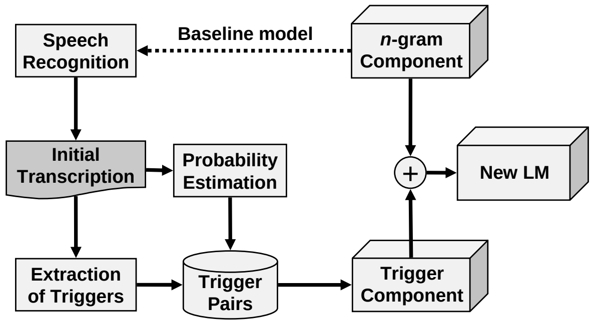
Description of Task and Corpora

- Task: NHK's *Sunday Discussion*
  - 1 hour panel discussions about political, economic issues
  - 10 programs chosen to cover diverse topics and sufficient variety of speakers
  - Recorded from June 2001 to January 2002
  - Average no. of utterances: 550 (14K words)
- Large corpus: National Diet (Congress) of Japan
  - Selected because of similarity in topic with Sunday Discussion
  - Recorded from 1999 to 2002
  - Total no. of documents: 2866 (71M words)
  - Documents for matched portion: 671 from year 2001 (17M words)

Proposed Approach

- Construct task-dependent trigger pairs from initial speech recognition results (initial transcription)
  - Homogeneous topics ⇒ Related keywords throughout sessions
  - Initial transcription erroneous but provides task-dependent info
- Problems
  - Small size of initial transcription
    - Insufficient to get enough trigger pairs and reliable estimates
  - Errors in initial transcription
    - Erroneous pairs increase probabilities of wrong words
- Solutions
  - Extract keywords with TF/IDF from whole discussion
    - Boost number of triggers and capture global constraints
  - Back-off scheme with statistics from large corpus
  - Use filtering techniques to discard unreliable pairs

Trigger-Based Adaptation from Initial Transcription



Construction of Trigger Pairs

- Extracted from *K*-best of initial transcription using term frequency/inverse document frequency (TF/IDF)

$$v_{ik} = \frac{tf_{ik}\log(N/df_k)}{\sqrt{\sum_{j=1}^{T}(tf_{ij})^2\left[\log(N/df_j)\right]^2}},$$

tf_{ik} ≡ Occurrence frequency of t_k in D_i
N ≡ Number of documents
df_k ≡ Number of documents containing t_k
T ≡ Number of terms in D_i

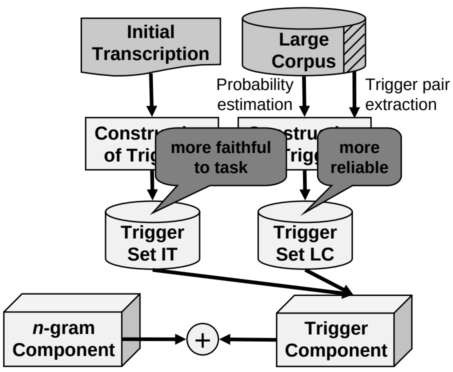
  - Create pairs from words with TF/IDF value greater than threshold
  - Only one document ⇒ IDF from same year portion of large corpus
- Probability estimated from *K*-best of initial transcription
  - Use text window of the previous *L* words
  - Probability of $w_1 \rightarrow w_2$ calculated as follows:

$$P_{TP}(w_2 \mid w_1) = \frac{N(w_1, w_2)}{\sum_j N(w_1, w_j)},\ N(\cdot,\cdot) \equiv \text{Co-occurrence frequency}$$

Filtering of Trigger Pairs

- To retain only topic words
  - POS-based filtering to remove function words
  - Stop word list filtering
    - List of most frequent words to be ignored
- To minimize incorrect trigger pairs
  - Confidence score filtering
    - Eliminate trigger pairs whose words have confidence score lower than threshold
  - Large corpus filtering
    - Extract trigger pairs also from large corpus and remove trigger pairs that are not in intersection

Back-off Scheme



Back-off Model

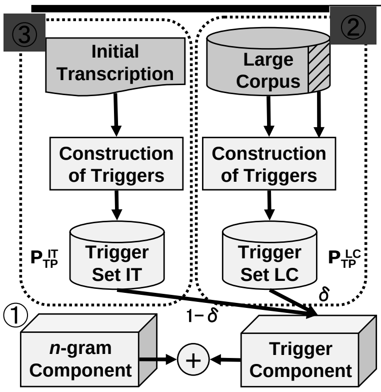
- Back off to trigger set LC when trigger pairs not found in set IT

$$P_{LM}(w_i \mid w_{i-L}^{i-1}) = \frac{1}{L}\sum_{j=i-L}^{i-1} P_{LM}(w_i \mid w_j)$$

$$P_{LM}(w_i \mid w_j) = \begin{cases} P_{NG}(w_i \mid w_{i-n+1}^{i-1}), \text{ if } P_{TP}^{IT}(w_k \mid w_j) = 0, P_{TP}^{LC}(w_l \mid w_j) = 0, \forall k, l \\ \lambda P_{NG}(w_i \mid w_{i-n+1}^{i-1}) + (1-\lambda)P_{TP}^{LC}(w_i \mid w_j), \text{ if } P_{TP}^{IT}(w_j \mid w_j) = 0, \forall j \\ \lambda P_{NG}(w_i \mid w_{i-n+1}^{i-1}) + (1-\lambda)\left(\delta P_{TP}^{LC}(w_i \mid w_j) + (1-\delta)P_{TP}^{IT}(w_i \mid w_j)\right), \text{ otherwise} \end{cases}$$

P_{NG} ≡ n-gram probability
P_{TP}^{IT} ≡ Probability of trigger set IT
P_{TP}^{LC} ≡ Probability of trigger set LC
λ ≡ Language model interpolation weight
δ ≡ Trigger set interpolation weight

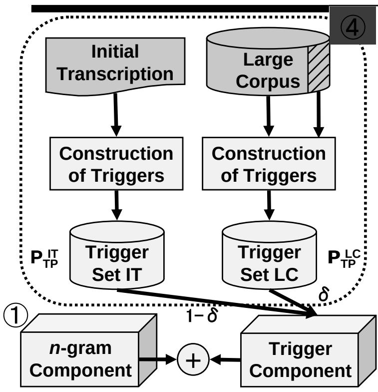
Experiments



① Baseline
② Large corpus (reference)
- δ = 1

③ Initial transcription (proposed)
- δ = 0

Experiments

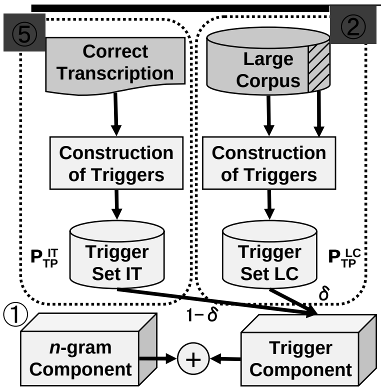


① Baseline
② Large corpus (reference)
- δ = 1

③ Initial transcription (proposed)
- δ = 0

④ Back-off Model (proposed)

Experiments



① Baseline
② Large corpus (reference)
- δ = 1

③ Initial transcription (proposed)
- δ = 0

④ Back-off Model (proposed)
⑤ Correct transcript (reference)

Experimental Setup

Task	Sunday Discussion 10 data sets (10 shows)
ASR system	Julius 3.5-rc2
Baseline LM	CSJ + National Diet trigram Linear interpolation (λ = 0.5)
Acoustic model	Triphone HMM from CSJ
Vocabulary	30K words
Out of vocabulary rate	1.56%
Baseline word accuracy	55.2%
Baseline perplexity	150

Perplexity Evaluation

Model	# pairs	Hit rate	PPL	Reduction (%)
① Baseline trigram	–	–	150	–
② Large corpus (LC)	9M	33%	121	19.33
③ Initial transcription (IT)	128K	31%	104	30.66
④ Back-off (IT+LC)	9M	35%	102	32.00
⑤ Correct transcription	71K	35%	73	51.33

- Reduction by IT much greater than that by LC
  - Effectiveness of proposed approach proved

Perplexity Evaluation

Model	# pairs	Hit rate	PPL	Reduction (%)
① Baseline trigram	–	–	150	–
② Large corpus (LC)	9M	33%	121	19.33
③ Initial transcription (IT)	128K	31%	104	30.66
④ Back-off (IT+LC)	9M	35%	102	32.00
⑤ Correct transcription	71K	35%	73	51.33

- The back-off model improved PPL slightly
  - The initial transcription provides well adapted trigger pairs ⇒ Benefit from LC is minimal
  - Efficacy with smaller initial transcriptions

Perplexity Evaluation

Model	# pairs	Hit rate	PPL	Reduction (%)
① Baseline trigram	–	–	150	–
② Large corpus (LC)	9M	33%	121	19.33
③ Initial transcription (IT)	128K	31%	104	30.66
④ Back-off (IT+LC)	9M	35%	102	32.00
⑤ Correct transcription	71K	35%	73	51.33

- Reduction by IT less than that by correct transcription
  - Half of the initial transcription has errors
  - ⇒ Results consistent with this fact

Self-triggers VS. Non-self-triggers

Model	# used pairs	PPL	Reduction (%)
Baseline trigram	–	150	–
Initial transcription (IT)	26K	104	30.66
Only self-triggers from IT	606	141	6.00
Only non-self-triggers from IT	26K	105	30.00

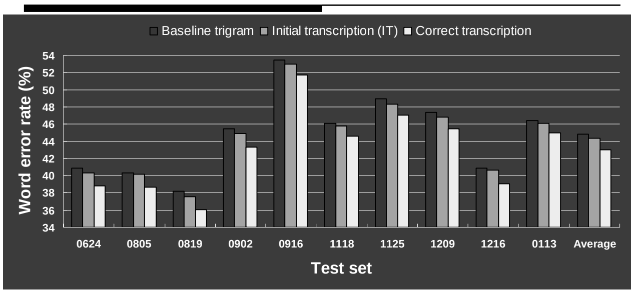
- Most perplexity reduction from non-self-triggers
  - Opposite to common finding in conventional trigger-based LM
  - Trigger pairs from IT are task-dependent and make a better match

n-gram Adaptation

- Create *n*-gram LM with *J*-best hypotheses
- Interpolate with baseline ⇒ adapted *n*-gram
- Interpolate with proposed trigger-based LM

Model	PPL	Reduction (%)
Baseline trigram	150	–
Adapted trigram	119	20.66
+ Initial transcription (IT)	87	42.00
+ Back-off model (IT+LC)	84	44.00

Speech Recognition Evaluation



- 0.98% relative improvement in WER by IT
- p-value = 0.022 ⇒ Statistically significant
- 4.07% relative improvement by correct transcription

Analysis of Results

- WER reduction << PPL reduction
  - Compared distributions of total extracted pairs and those used during PPL and WER evaluation
    - Trigger pairs not found in correct transcription are labeled as incorrect

	Class of triggers	Entries	Count	Proportion	
Total pairs	Correct	31253	–	24.23	–
	Incorrect	97727	–	75.77	–
Pairs used in PPL	Correct	14848	26716	97.37	98.36
	Incorrect	401	446	2.63	1.64
Pairs used in WER	Correct	7441	30290	43.91	52.88
	Incorrect	9505	26987	56.09	47.12

Summary

- Novel trigger-based LM adaptation using initial transcription and large corpus
- Remarkable improvement in PPL over baseline and typical trigger-based LM
- Most improvement from non-self-triggers
- Further improvement by *n*-gram adaptation
- Extracted trigger pairs are task-dependent