Discriminative Rescoring using Landmarks

Katrin Kirchhoff

Rationale

- WS04 approach: lattice/N-best list rescoring instead of first-pass recognition
- baseline system already provides high-quality hypotheses
  - 1-best error rate from N-best lists: 24.4% (RT-03 dev set)
  - oracle error rate: 16.2%
- $\Rightarrow$ use landmark detection only where necessary, to correct errors made by baseline recognition system

Example

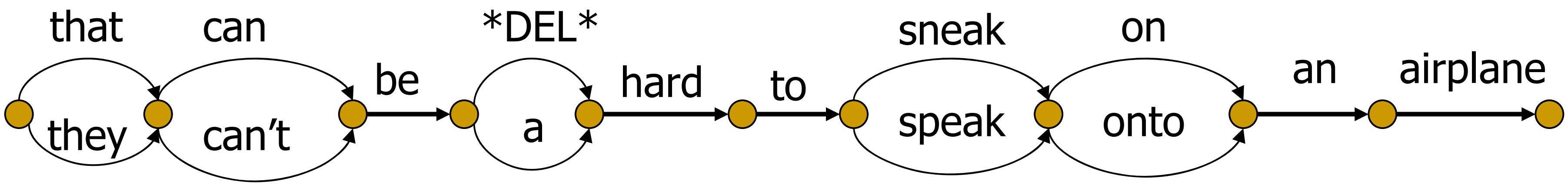
fsh_60386_1_0105420_0108380

Ref: that cannot be   that hard to sneak onto an airplane
Hyp: they can   be a that hard to speak on   an airplane

- Identify word confusions
- Determine most important acoustic-phonetic features that distinguish confusable words
- Use high-accuracy landmark detectors to determine probability of those features
- Use resulting output for rescoring

Identifying Confusable Hypotheses

- Use existing alignment algorithms for converting lattices into confusion networks (Mangu, Brill & Stolcke 2000)



- Hypotheses ranked by posterior probability
- Generated from n-best lists without 4-gram or pronunciation model scores ($\Rightarrow$ higher WER compared to lattices)
- Multi-words ("I_don't_know") were split prior to generating confusion networks

Identifying Confusable Hypotheses

- How much can be gained from fixing confusions?
- Baseline error rate: 25.8%
- Oracle error rates when selecting correct word from confusion set:

# hypotheses to select from	Including homophones	Not including homophones
2	23.9%	23.9%
3	23.0%	23.0%
4	22.4%	22.5%
5	22.0%	22.1%

Selecting relevant landmarks

- Not all landmarks are equally relevant for distinguishing between competing word hypotheses (e.g. vowel features irrelevant for *sneak* vs. *speak*)
- Using all available landmarks might deteriorate performance when irrelevant landmarks have weak scores (but: redundancy might be useful)
- Automatic selection algorithm
  - Should optimally distinguish set of confusable words (discriminative)
  - Should rank landmark features according to their relevance for distinguishing words (i.e. output should be interpretable in phonetic terms)
  - Should be extendable to features beyond landmarks

Selecting relevant landmarks

- Words are associated with variable-length sequences of landmarks
- Options for selection:
  - Use a discriminative sequence model: Conditional Random Fields
  - Convert words to fixed-length representation and use standard discriminative classifier, e.g. maximum-entropy model, MLP, SVM
  - Related work (e.g. by Byrne, Gales): Fisher score spaces + SVMs
  - Here: phonetic vector space + maxent model (interpretable)

Maximum-Entropy Landmark Selection

- Convert each word in confusion set into fixed-length landmark-based representation using idea from information retrieval:
- Vector space consisting of binary relations between two landmarks
  - Manner landmarks: precedence, e.g. *V < Son. Cons.*
  - Manner & place features: overlap, e.g. *V o +high*
  - preserves basic temporal information
- Words represented as frequency entries in feature vector
- Not all possible relations are used (phonotactic constraints, place features detected dependent on manner landmarks)
- Dimensionality of feature space: 40 - 60
- Word entries derived from phone representation plus pronunciation rules

Vector-space word representation

	Start < Fric	Fric< Stop	Fric< Son	Fric < Vowel	Stop < Vowel	Vowel o high	Vowel o front	Fric o strident
speak	1	1	0	0	1	1	1	1
sneak	1	0	1	0	0	1	1	1
seek	1	0	0	1	0	1	1	1
he	1	0	0	1	0	1	1	0
she	1	0	0	1	0	1	1	1
steak	1	1	0	0	1	0	1	1
….								

Maximum-entropy discrimination

- Use maxent classifier

$$P(y \mid x) = \frac{1}{Z(x)} \exp\left(\sum_k \lambda_k f_k(x, y)\right)$$

- Here: *y* = words, *x* = acoustics, *f* = landmark relationships
- Why maxent classifier?
  - Discriminative classifier
  - Possibly large set of confusable words
  - Later addition of non-binary features
- Training: ideally on real landmark detection output
- Here: on entries from lexicon (includes pronunciation variants)

Maximum-entropy discrimination

- Example: sneak vs. speak

sneak		speak	
SC ○ +blade	2.47	SC ○ +blade	-2.47
FR < SC	2.47	FR < SC	-2.47
FR < SIL	-2.11	FR < SIL	2.11
SIL < ST	-1.75	SIL < ST	1.75
…..		…..	

- Different model is trained for each confusion set $\Rightarrow$ landmarks can have different weights in different contexts

Landmark queries

- Select N landmarks with highest weights
- Could scan bottom-up landmark detection output for presence of relevant landmarks
- Better: use knowledge of relevant landmarks in top-down fashion (suggestion by Jim)
- Ask landmark detection module to produce scores for selected landmarks within word boundaries given by baseline system
- Example:

Confusion networks

sneak 1.70 1.99 SC ○ +blade ?

sneak 1.70 1.99 SC ○ +blade 0.75 0.56

Landmark detectors

Rescoring

- Landmark detection scores: weighted combination of manner and place probabilities
- Normalization across words confusion set & combination (weighted sum or product) with original probability distribution given by baseline system
- Or: use as additional features in a maxent model for rescoring confusion networks (more on this in Kemal's talk)
- Only applied to confusion sets that contain phonetically distinguishable hypotheses (e.g. not *by* - *buy*, *to*-*two*-*too*…)
- Only applied to sets where words do not compete with DELETE

Experiments

- Varied number of landmark scores to use (1,2,…all)
- Top 2 vs. 3 vs. all hypotheses in confusion network
- Use of entire word time interval vs. restricting time intervals to approximate location of landmarks
- Changes in feature-space representation of lexicon
- Various score combination methods for rescoring
- Initial experiments on learning lexicon representation from data (for most frequent words)

Results

RT-03 dev set, 35497 Words, 2930 Segments, 36 Speakers (Switchboard and Fisher data)

	WER	Insertions	Deletions	Substitutions
Baseline	25.8%	2.6% (982)	9.2% (3526)	14.1% (5417)
Rescored	25.8%	2.6% (984)	9.2% (3524)	14.1% (5408)

Rescored: product combination of old and new prob. distributions, weights 0.8 (old), 0.2 (new)

-Correct/incorrect decision changed in about 8% of all cases
-Slightly higher number of fixed errors vs. new errors

Analysis

- When does it work?
  - Detectors give high probability for correct distinguishing feature

    mean (correct) vs. *me* (false)   V < +nasal 0.76

- When does it not work?
  - Problems in lexicon representation

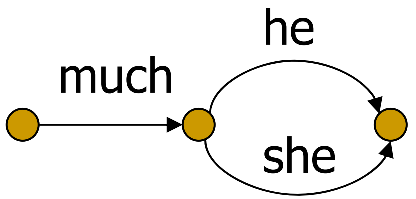
    once (correct) vs. *what* (false):   Sil ○ +blade 0.87

    *can't [*kæ̃t] (correct) vs *cat* (false):   SC ○ +nasal 0.26
  - Landmark detectors are confident but wrong

    like (correct) vs. *liked* (false): Sil ○ +blade 0.95

Analysis

- Incorrect landmark scores often due to word boundary effects, e.g.:



- Word boundaries given by baseline system may exclude relevant landmarks or include parts of neighbouring words

Conclusions

- Positive trend but not strong enough yet to decrease word error rate
- Method can be used with classifiers other than landmark detectors (e.g. high-accuracy triphone classifiers)
- Can serve as diagnostic tool (statistics of score queries $\Rightarrow$ relevance of phonetic distinctions for improving word error rate on given corpus)
- Provides information about which detector outputs are likely to help vs. likely to cause errors $\Rightarrow$ feedback for developing landmark classifiers
- Advantage: little computational effort, fast

Future Directions

- Improve landmark detectors (e.g. specialized detectors for word endings)
- Select landmarks that are not only discriminative but can also be detected robustly
- Learn lexical representation from data (takes into account errors made by detectors)
- Change lexical representation to include more temporal constraints
- Try approach with other classifiers
- Allow flexible word segmentation