Book Chapter 5

Peter A. Heeman, James F. Allen: Improving Robustness by Modeling Spontaneous Speech Events

Presented by: Teemu Hirsimäki

Spontaneous speech events

- Especially in human-to-human speech, people tend to group words into intonational phrases and make repairs.

  um it'll be there
  it'll get to Dansville at three a.m.
  and then you wanna
  do you take tho-
  want to take those back to Elmira

- This causes problems for traditional language models.

Purpose of the paper

- Traditionally spontaneous events are considered as noise.
- Here we also try to model:
  - Part-of-speech (POS) tagging (verbs, prepositions, nouns)
  - Intonational phrases
  - Editing terms (*um, let's see*)
  - Speech repairs

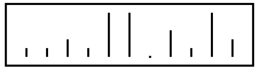
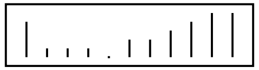
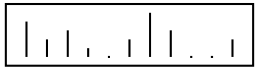
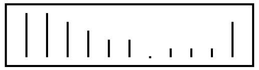
Speech repairs

Fresh start:	*I need to send* reparandum	×	*let's see* editing term	*how many do you need?* alteration
Modification repair:	*I need one* reparandum	×	*um* editing term	*two boxcars.* alteration
Abridged repair:	*We need to*	×	*um* editing term	*get the bananas.*

× = interruption point

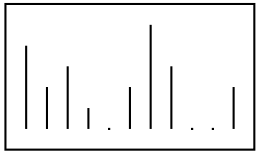
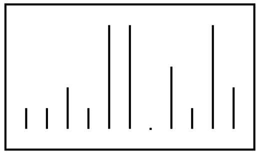
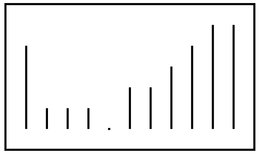
Estimating word probabilities

- Estimating a word distribution for all possible contexts is naturally impossible:

contexts	**probability of the next word**
on the street, there was a	
the presentation about spontaneous speech events was	
the book on	
the power consumption of	
early on the	
⋮	⋮

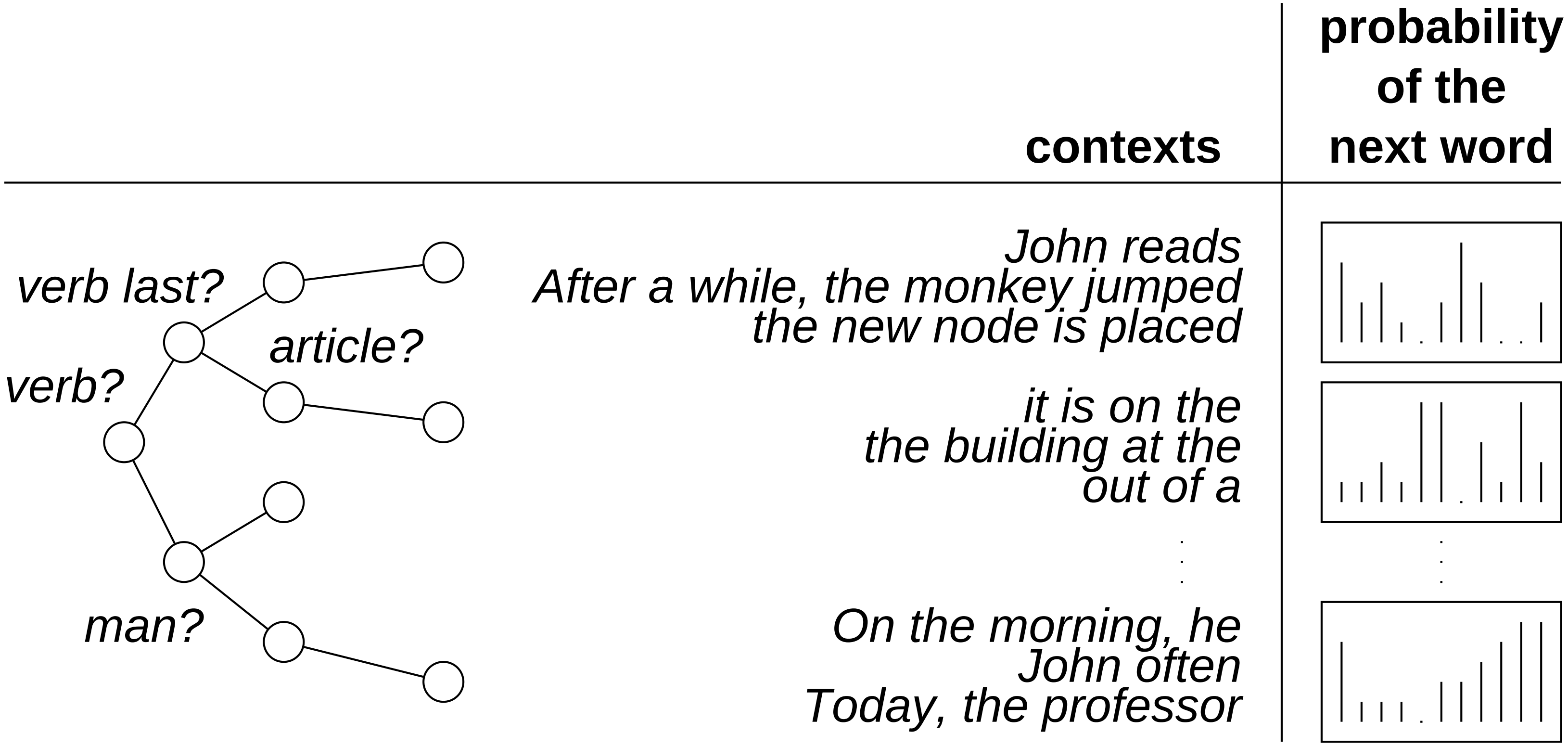
Trigram model

- The trigram model discards all but the last few words:

contexts	probability of the next word
on the street, there was a *John was a* *the course was a*	
the book on the *the car on the* *see the figure on the*	
⋮	⋮
it was cheap , but *would improve the algorithm , but* *sun shines today , but*	

Decision tree

- The contexts can also be clustered using a decision tree:



contexts

probability of the next word

John reads
After a while, the monkey jumped
the new node is placed

it is on the
the building at the
out of a

⋮

On the morning, he
John often
Today, the professor

Modeling part-of-speech tags

- Traditionally: $\hat{W} = \mathrm{argmax}_W \, P(W|A)$
- First the POS tags are included in the model:

$$\begin{aligned} \hat{W}\hat{P} &= \operatorname*{argmax}_{WP} P(WP|A) \\ &= \operatorname*{argmax}_{WP} P(A|WP)P(WP) \end{aligned}$$

Identifying repairs and intonational phrases

- In addition to W_i and P_i, three new variables are introduced:
  - Repair variable R_i = { *MOD*, *CAN*, *ABR*, *NULL* }
  - Editing term variable E_i = { *PUSH*, *ET*, *POP*, *NULL* }
  - Intonation variable I_i = { *%*, *NULL* }
- *that's the one with the bananas % PUSH I ET mean POP MOD that's taking the bananas*
- The speech recognition problem becomes:

$$\hat{W}\hat{P}\hat{R}\hat{E}\hat{I} = \operatorname*{argmax}_{WPREI} P(WPREI|A)$$

- An example:

  it takes one **PUSH** *you* **ET** *know* **POP** **MOD** *two hours* **%**

- The following contexts are given to decision tree:
  - *it*-**PRP** *takes*-**VBP** *one*-**CD** **PUSH** ⇐ *you*
  - *it*-**PRP** *takes*-**VBP** *one*-**CD** **PUSH** *you*-**PRP** *know*-**VBP** ⇐ **POP**
  - *it*-**PRP** *takes*-**VBP** *one*-**CD** ⇐ **MOD**

    (also with editing term)
  - *it*-**PRP** *takes*-**VBP** *one*-**CD** **MOD** ⇐ *two*

Correcting repairs

- Three additional variables are defined:
  - Reparandum onset $O_{ij} = \{$ ONSET, NULL $\}$
  - Correspondence licensor $L_{ij} = \{$ CORR, NULL $\}$
  - Word Correspondence $C_i = \{$ M, R, X, NULL $\}$
- *you can $\underline{\textit{carry them both}}$ % $\underline{\textit{bring both}}$ here*
- The recognition problem becomes:

$$\begin{aligned}\hat{W}\hat{P}\hat{C}\hat{L}\hat{O}\hat{R}\hat{E}\hat{I} &= \underset{WPCLOREI}{\operatorname{argmax}} \; P(WPCLOREI|A) \\ &= \underset{WPCLOREI}{\operatorname{argmax}} \; P(A|WPCLOREI)P(WPCLOREI)\end{aligned}$$

Trains corpus

- 6.5 hours of speech
- 34 different speakers
- Transcription
- POS tags, discourse markers, end-of-turns
- Intonational phrase boundaries

Experiments: POS tagging

	WP	WPCLOREI	WPCLOREIS
POS errors	1711	1652	1563
POS error rate	2.93	2.83	2.68
Word perplexity	24.04	22.96	22.35

The rightmost model uses the amount of silence to adjust the probability distributions of repair, editing term and intonation variables.

Experiments: Intonational phrases

Type	Recall	Precision	Error rate
Within turn	70.76	70.82	57.79
End of turn	98.05	94.17	8.00
All boundaries	84.76	82.53	33.17

- Recall: correct identifications over all events
- Precision: correct identifications over all identifications
- Error rate: number of errors over all events

Experiments: Repair detection

Type	Recall	Precision	Error rate
All repairs	76.79	86.66	35.01
Abridged	75.88	82.51	40.18
Modification	80.87	83.37	35.25
Fresh starts	48.58	69.21	73.02
Mod. & Fresh	73.69	83.85	40.49

Experiments: Repair correction

Type	Recall	Precision	Error rate
All repairs	65.85	74.32	56.88
Abridged	75.65	82.26	40.66
Modification	77.95	80.36	41.09
Fresh starts	36.21	51.59	97.76
Mod. & Fresh	63.76	72.54	60.36

Conclusions

- The new model into account part-of-speech tags, intonational phrases and speech repairs.
- Benefits of the model are:
  - Identification of intonational phrases
  - Detection and correction of speech repairs
  - Richer output for later processing
- Decision tree algorithm was used to train the complex probability distribution.
- Improvements in POS tagging and perplexity results.
- In future, more acoustic cues should be used.