Role Recognition

Ke-Yu Chen
01/25/2010

References

Mainly focus on

1. R. Barzilay, M. Colins, J. Hirschberg, and S. Whittaker. 2000. **The rules behind roles: Identifying speaker role in radio broadcasts**. Proc. AAAI Conference on Artificial Intelligence & Conference on Innovative Applications of Artificial Intelligence, 679-684. AAAI Press/MIT Press.
2. N. Garg, S. Favre, H. Salamin, D. Hakkani-Tur, and A. Vinciarelli. 2008. **Role recognition for meeting participants:an approach based on lexical information and social network analysis**. Proceedings ACM International Conference on Multimedia, 693-696.

Others

3. Singer, Y., and Shapire, R. 1998. **Improved boosting algorithms using confidence-rated predictions**. In Proceeding of 11th Annual Conference on Computational Learning Theory, 80–91.
4. **Beam search** :http://en.wikipedia.org/wiki/Beam_search
5. Mencher, M. 1987. **News Reporting and Writing**. Dubuque, Iowa: William C. Brown, 4 edition.
6. H. Tischler. **Introduction to Sociology**. Harcourt Brace College Publishers, 1990.
7. I. McCowan, J. Carletta, W. Kraaij, S. Ashby, S. Bourban, M. Flynn, M. Guillemot, T. Hain, J. Kadlec, V. Karaiskos, M. Kronenthal, G. Lathoud, M. Lincoln, A. Lisowska, W. Post, D. Reidsma, and P. Wellner. **The ami meeting corpus**. In Proceedings of the 5th International Conference on Methods and Techniques in Behavioral Research, 2005.
8. S. Wasserman and K. Faust. **Social Network Analysis: Methods and Applications**. Cambridge University Press, 1994.
9. J. Dines, et al. **The segmentation of multi-channel meeting recordings for automatic speech recognition**. In Proceedings of the Interspeech, pages 1213–1216, 2006.

The Rules Behind Roles: Indentifying Speaker Role in Radio Broadcasts [1]

Motivation

- Providing information about story structure is critical for browsing audio broadcasts

- Speaker role is an important cue to story structure

Goals

- From broadcast news programs, identify
  - Anchor
    - Reading news
    - Introducing reporters from journalists
    - Announce upcoming events

  - Journalist
    - Professional speakers (usually in remote locations)
    - Interview with guests
  - Guest speaker
    - Non-professional speakers addressing a subjective point of view

Features used in role identification

- Lexical features
- Features from surrounding context
- Duration of a segment
- Explicit speaker introductions

Lexical features

- Signature phrases
  - “*This is CNN’s Prime news*”
  - Frequently used by anchor and journalist
- Planned vs. spontaneous speech
  - “*Well, you know…*” more likely used by guest speakers
- Capitalization
  - The word “*Clinton*” tends to be capitalized

Features from surrounding context

- Label and content of adjacent segments may predict current speaker type

- Individual stories are usually
  - Started by an anchor introduction
  - A journalist introduction
  - Alternation between journalist and guest segments

- But, sometimes…
  - There is no guest speakers (video)
  - Talks may be initiated and dominated by a guest speaker… (video)

Duration of a segment

- Journalist guide books [5] advise controlling
  - Time length of guest speaker segments
  - Lengths for anchor lead-ins / journalist’s questions

Explicit speaker introductions

- Professional speakers usually need to introduce themselves or other speakers
  - "*This is Mike & Mike, ESPN*"
  - "*Thanks Claudio Sanchez for that report*"

- Indentify and tag words (i.e. *Mike & Mike* or *Claudio Sanchez*)

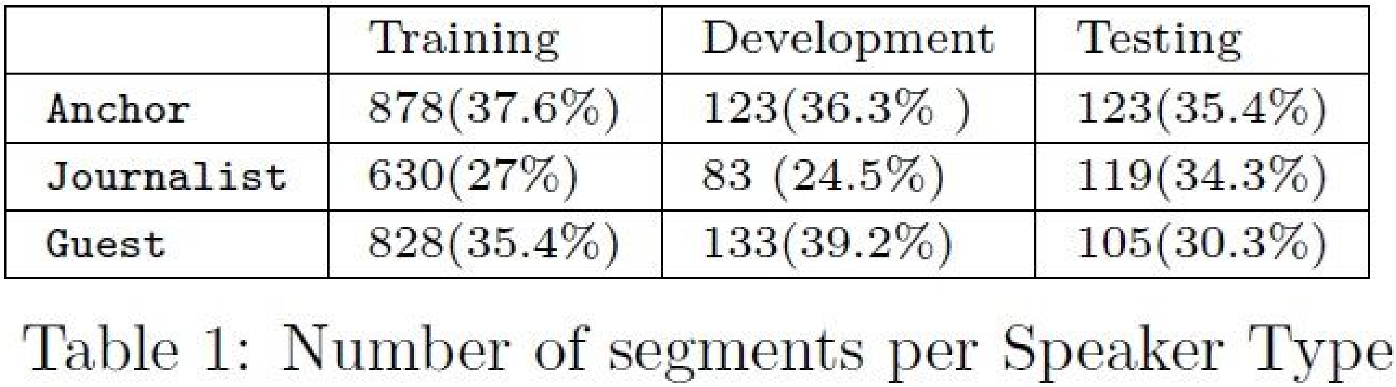
Experimental setups (1)

- Input: ASR transcriptions
  - NIST TREC SDR corpus (35.5 hr broadcast news)
  - Segmenting the speech into audio paragraphs
  - Produce the transcription using ASR

- Output: a label (one of the roles) with each segments

Experimental setups (2)

- Total 37 broadcasts
  - Training sets (27 broadcasts)
    - A set of segments with known labels to train a classifier
  - Development sets (5 broadcasts)
  - Held-out test set (5 broadcasts)



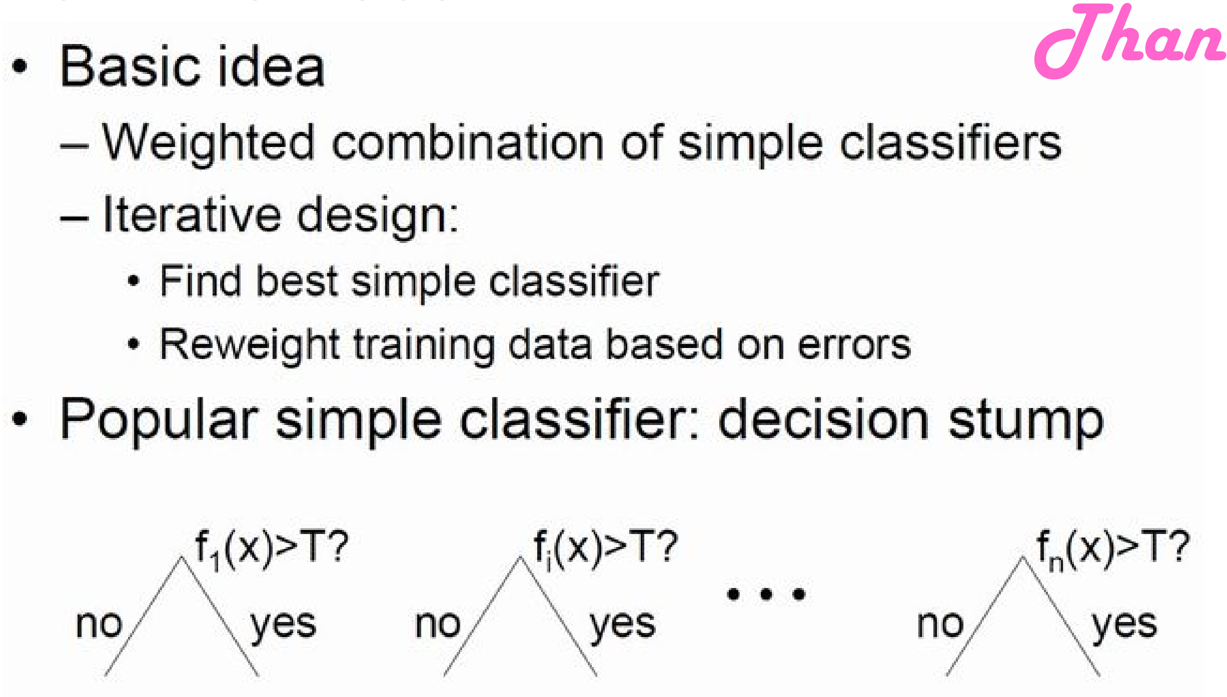
	Training	Development	Testing
Anchor	878(37.6%)	123(36.3%)	123(35.4%)
Journalist	630(27%)	83 (24.5%)	119(34.3%)
Guest	828(35.4%)	133(39.2%)	105(30.3%)

Table 1: Number of segments per Speaker Type

Learning methods (1)

- BoosTexter [3]
- Maximum entropy modeling

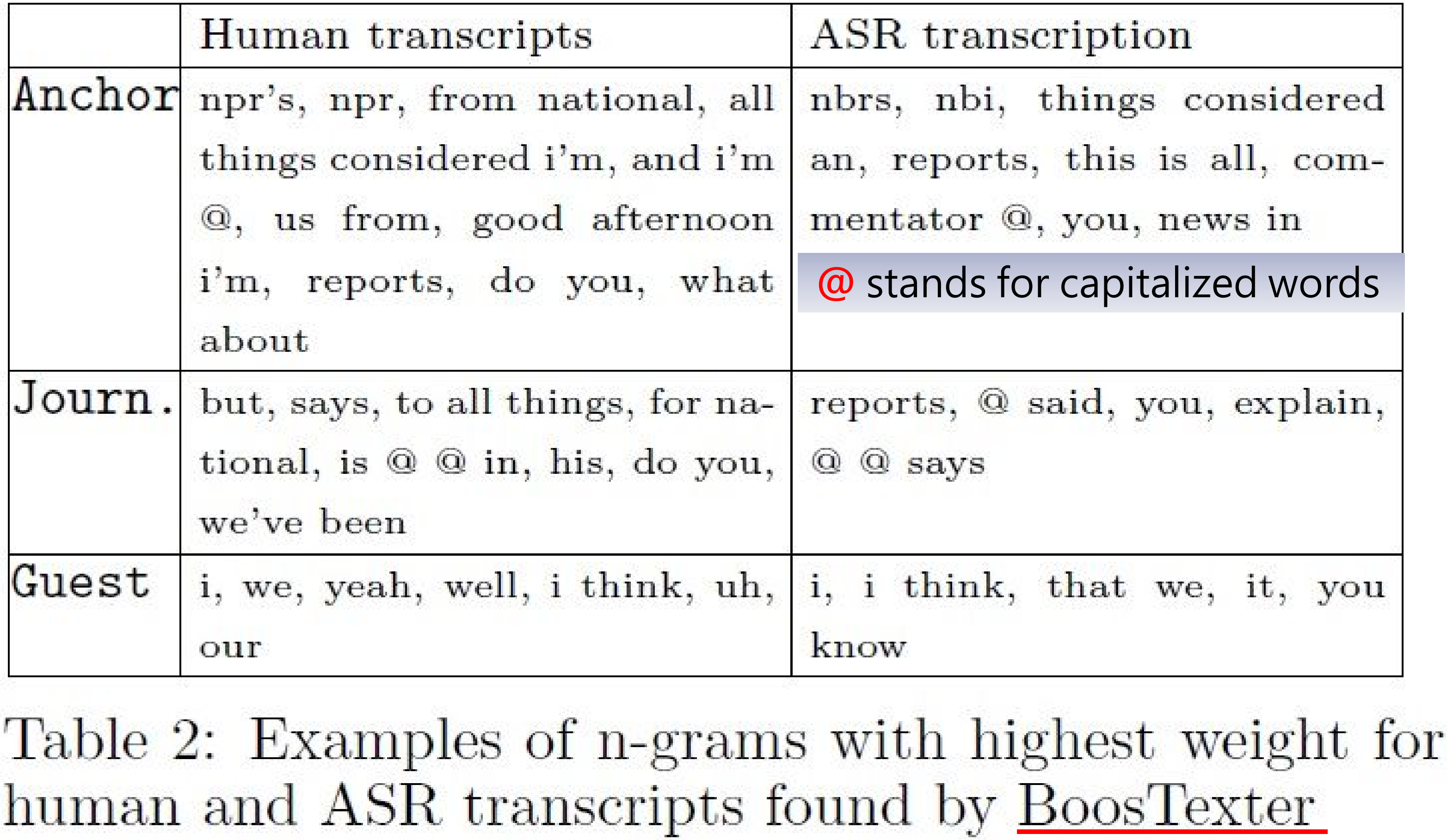
- Both methods



- Basic idea
  - – Weighted combination of simple classifiers
  - – Iterative design:
    - Find best simple classifier
    - Reweight training data based on errors
- Popular simple classifier: decision stump

Thank you, prof.

Learning methods (2)



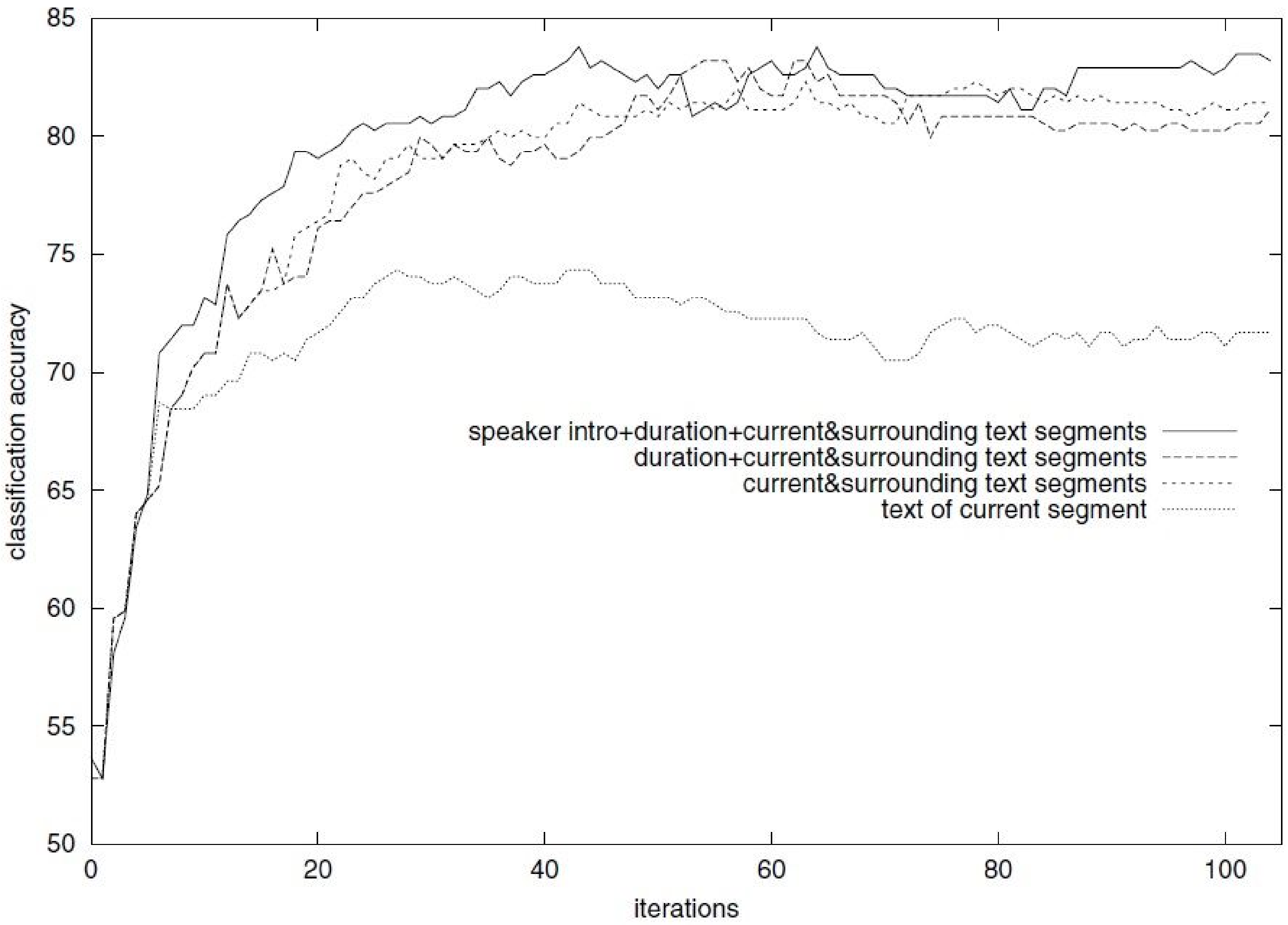
	Human transcripts	ASR transcription
Anchor	npr's, npr, from national, all things considered i'm, and i'm @, us from, good afternoon i'm, reports, do you, what about	nbrs, nbi, things considered an, reports, this is all, commentator @, you, news in
Journ.	but, says, to all things, for national, is @ @ in, his, do you, we've been	reports, @ said, you, explain, @ @ says
Guest	i, we, yeah, well, i think, uh, our	i, i think, that we, it, you know

@ stands for capitalized words

Table 2: Examples of n-grams with highest weight for human and ASR transcripts found by BoosTexter

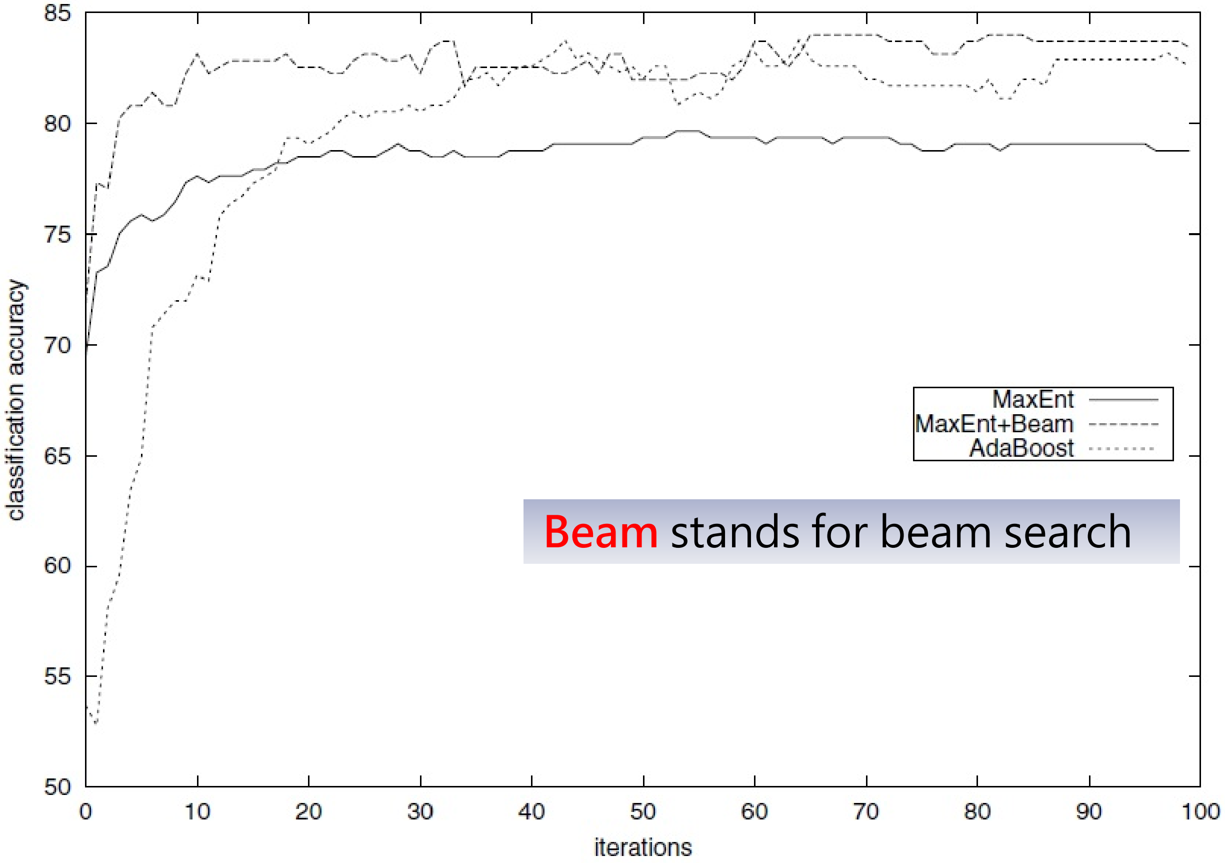
Results (1)

- Classification accuracy using different features



Results (2)

- Classification accuracy of different learning algorithms

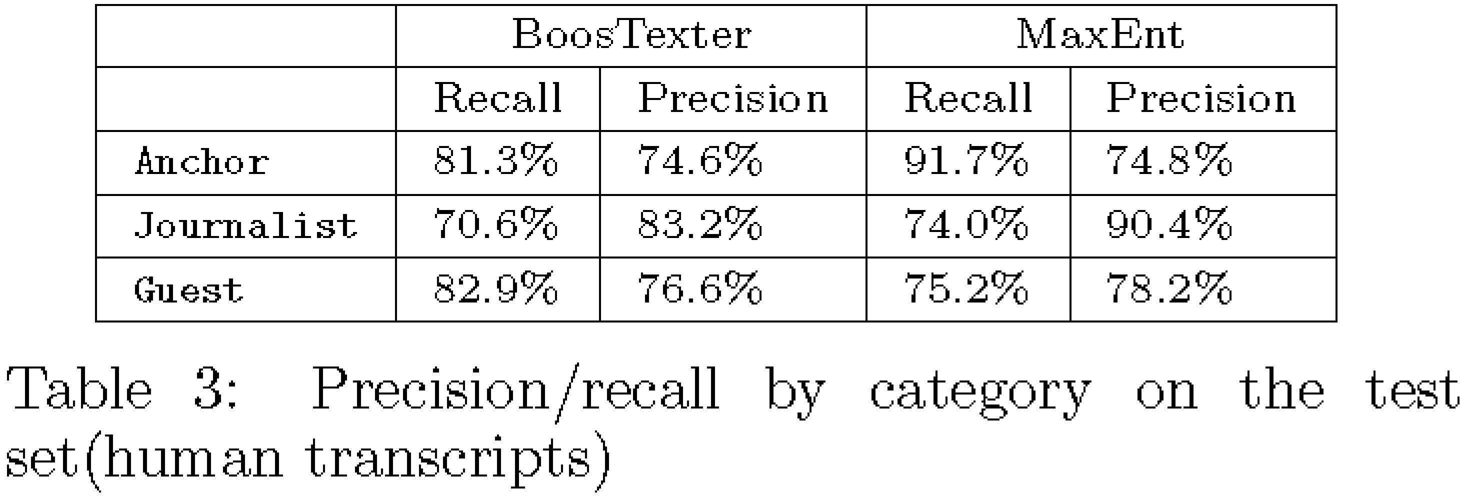


Beam stands for beam search

Results (3)

- Negative "chain reaction"
  - In BoosTexter, labels of 2 previous seg. were given
  - Drops accuracy

    sometimes, categories of previous speakers fully determine the category of current speaker (e.g. Anchor → Journalist)



	BoosTexter		MaxEnt	
	Recall	Precision	Recall	Precision
Anchor	81.3%	74.6%	91.7%	74.8%
Journalist	70.6%	83.2%	74.0%	90.4%
Guest	82.9%	76.6%	75.2%	78.2%

Table 3: Precision/recall by category on the test set(human transcripts)

Conclusion

- Exploits the lexical information (from ASR transcriptions) to identify 3 type of roles
  - Anchor, Journalist, Guest speakers

Role Recognition for Meeting Participants: an Approach Based on Lexical Information and Social Network Analysis [2]

Motivation

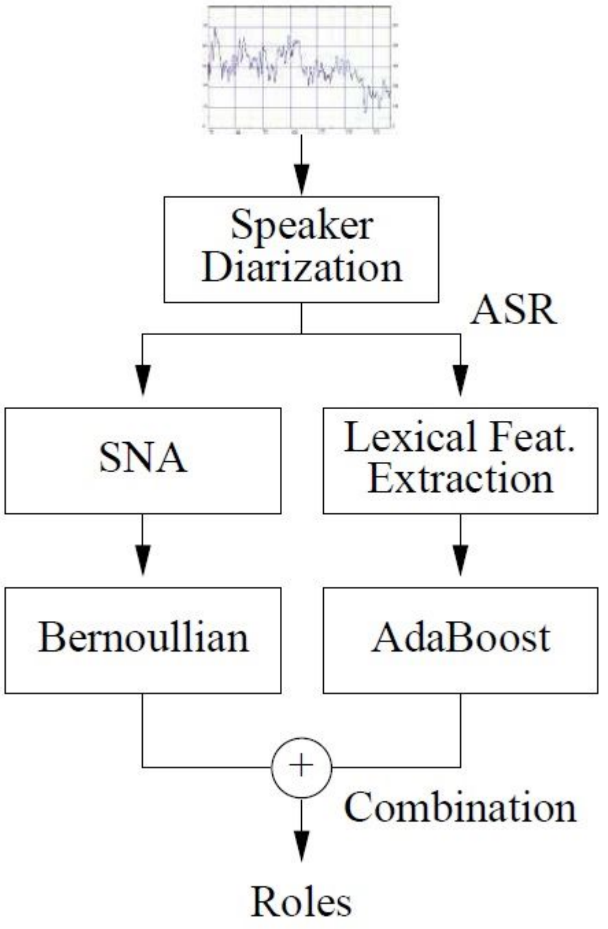
- "*People do not interact with one another as anonymous beings. They come together in the context of specific environments and with specific purposes.*" [6]

- In role recognition, consider not only lexical features but also the effect of social network

Goals

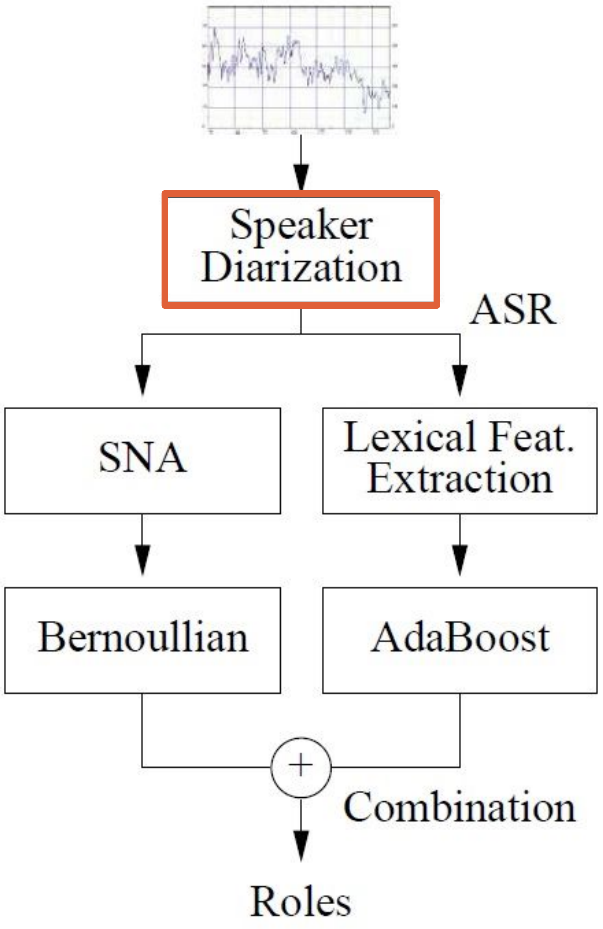
- In a meeting, identify
  - Project Manager (PM)
  - Marketing Expert (ME)
  - User Interface Expert (UI)
  - Industrial Designer (ID)

Approach – overview



- Combination of
  - Lexical features (right)
  - Social network (left)

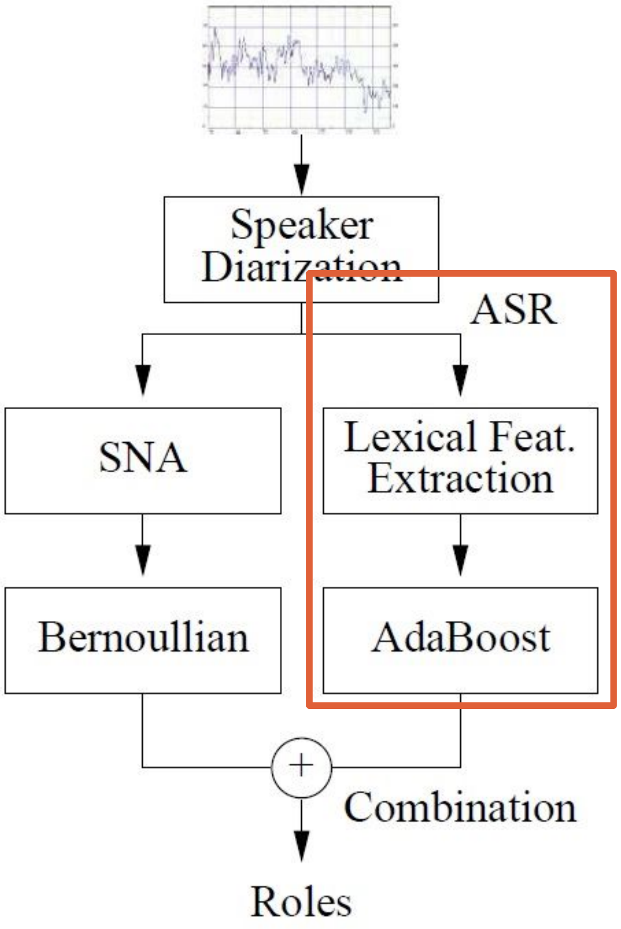
Approach – diarization [9]



- Indentify time intervals where each speaker talks

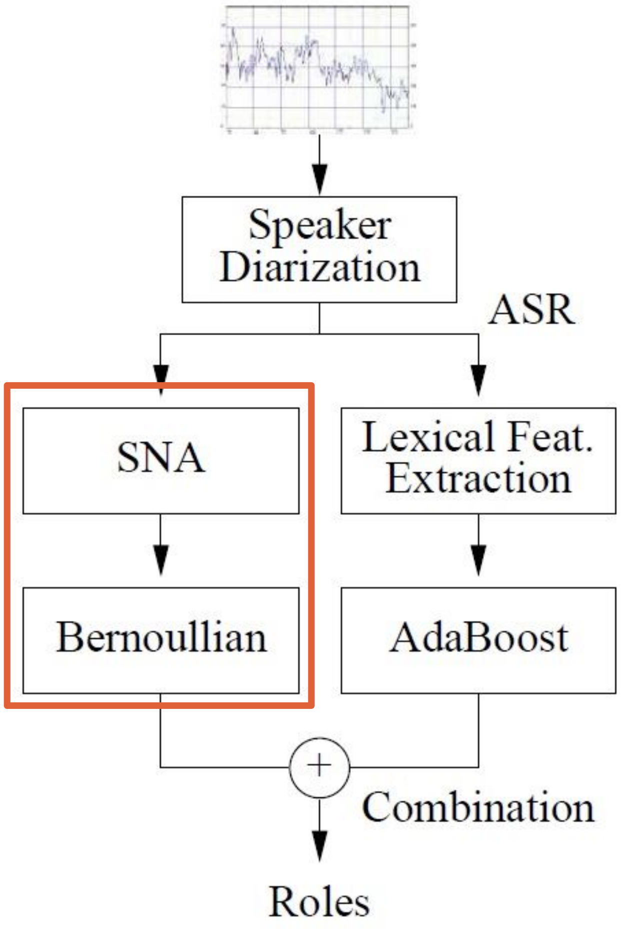
- Each meeting recording is divided into segments

Approach – lexicon based (right)



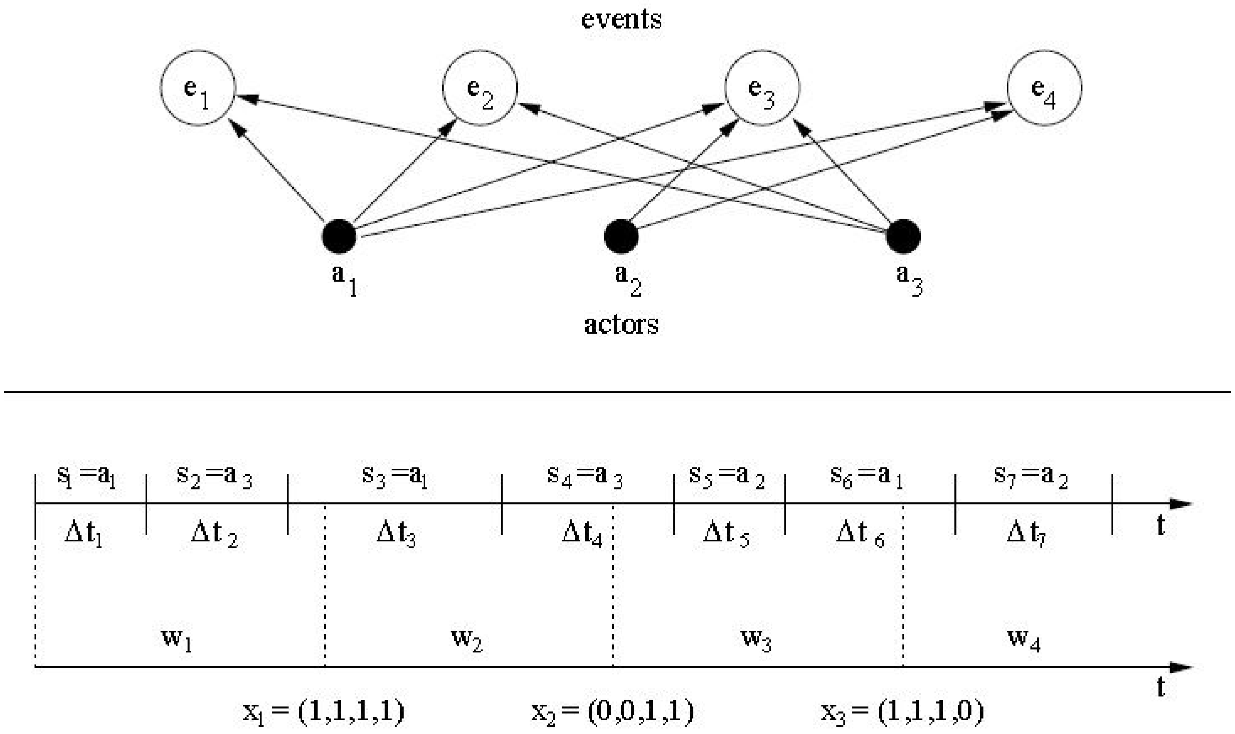
- Lexical features extraction from ASR transcripts
- Mapping features into roles (`BoosTexter`)
- ASR induces noises

Approach – SNA based (left)



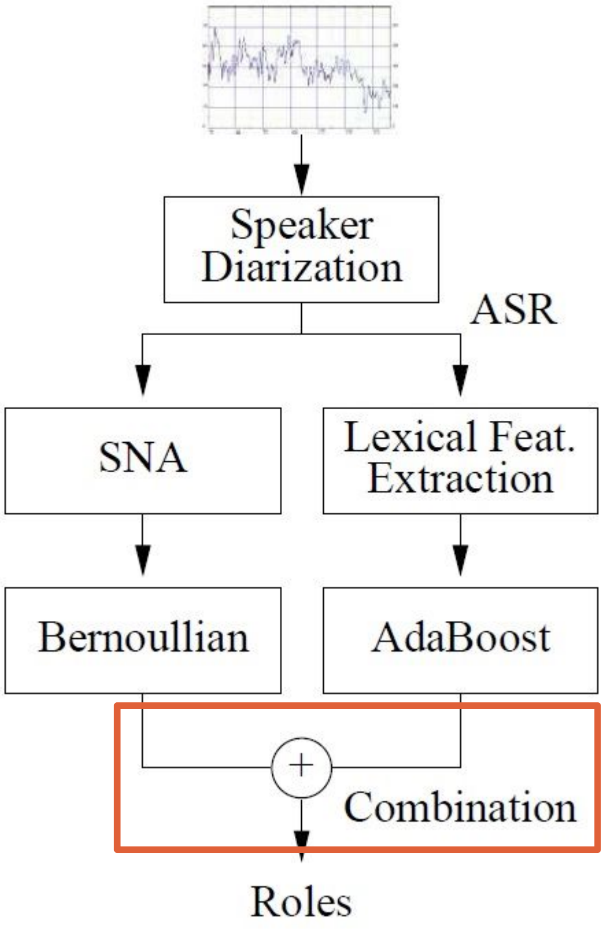
- Interactions between participants
- Bernoulli distribution

Social network analysis (SNA) [10]



- Actor nodes (a_i) and event nodes (e_i)
- Link = an actor participate with an event
- Uniform segments (W_i)

Approach – combination

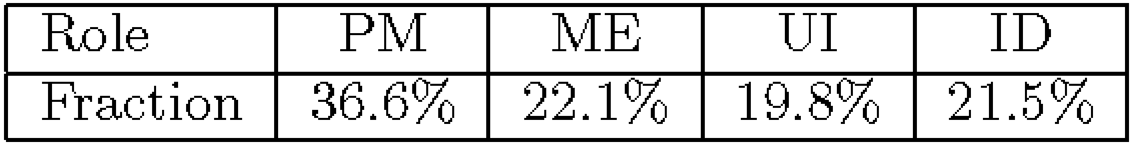


- Coefficient (b) is selected through cross validation (?)

$$
\begin{aligned}
r^* &= \arg\max_{r\in\mathcal{R}} p(\vec{x},\vec{d}\,|\,r,\vec{\mu}_r) \\
&= \arg\max_{r\in\mathcal{R}} \underline{\beta} \log p(\vec{d}\,|\,r) + \underline{(1-\beta)} \log p(\vec{x}\,|\,\vec{\mu}_r)
\end{aligned}
$$

Experimental setup

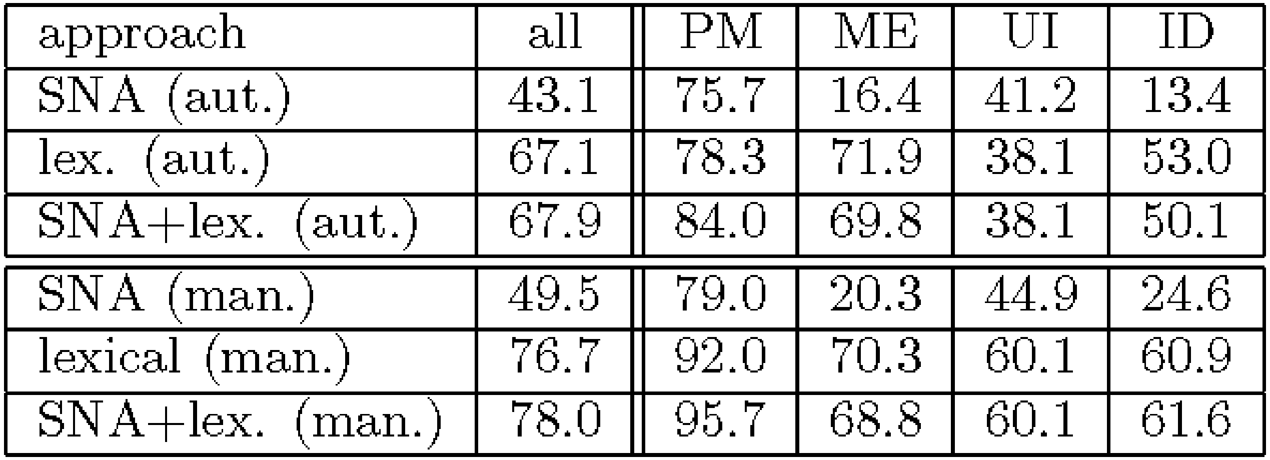
- AMI corpus [7] (138 meetings, 45.5 hr)
- Role distribution



Role	PM	ME	UI	ID
Fraction	36.6%	22.1%	19.8%	21.5%

- Leave-one-out
  - All meetings of the corpus are used for training except one that is left as the test set

Results



approach	all	PM	ME	UI	ID
SNA (aut.)	43.1	75.7	16.4	41.2	13.4
lex. (aut.)	67.1	78.3	71.9	38.1	53.0
SNA+lex. (aut.)	67.9	84.0	69.8	38.1	50.1
SNA (man.)	49.5	79.0	20.3	44.9	24.6
lexical (man.)	76.7	92.0	70.3	60.1	60.9
SNA+lex. (man.)	78.0	95.7	68.8	60.1	61.6

Groundtruth

- Lexical feature is more robust
- SNA does not perform well (43.1%)
  with ME & ID even lower than chance (25%)
  - SNA makes more sense when # of participant ↑

Conclusions

- Identify one of the four predefined roles for each segment in a meeting
- Combine lexical features and social network (SNA)
- Lexical features are more robust, while SNA tends to perform better when the number of participants increases

Discussions (1)

- Observations in [1]
  - Anchors tend to occur more frequently in the program
  - Guest segments never introduce a story
  - Speaker transition

    When a journalist stop speaking, it sometimes means a story has ended
  - Speaker change may imply story boundaries

    Acoustic characteristics of speakers (not used in this paper)

Discussion (2)

- How could role identification help us?
  - Enhance browser

    users can access specific data segments based on role
  - Summarization

    segments corresponding to certain roles can be retained in the summary since it is more representative (e.g. Anchor's introduction)
  - Thematic segmentation

    specific roles are related to specific topics

Discussion (3)

- Extract information from videos

- More background information (prior)
  - Indoor (anchor), outdoor(journalist/guest)
  - Light condition (bright vs. dark)
  - Background noise
  - Location or building
  - …

More…