An interactive timeline for Speech Database Browsing

Benoit Favre

SRI – STAR Lab Seminar Series
2007-08-02

Who am I?

- Benoit Favre
  - PhD "Automatic Speech Summarization", at LIA
  - Postdoc at ICSI until March 2008 (sentence segmentation)
  - favre@icsi.berkeley.edu
- Former lab: *Laboratoire Informatique d'Avignon* (LIA)
  - http://www.lia.univ-avignon.fr – English coming soon
  - Speech group (~10 permanent and 20 PhD students)
    - Dialogue systems (Renato De Mori)
    - Speaker id/diarization (Alize toolkit, Jean-François Bonastre)
    - STT: French and resource-sparse languages
    - Voice/Language pathologies

Who am I?

- Benoit Favre
  - PhD "Automatic Speech Summarization", at LIA
  - Postdoc at ICSI until March 2008 (sentence segmentation)
  - favre@icsi.berkeley.edu
- Former lab: *Laboratoire Informatique d'Avignon* (LIA)
  - http://www.lia.univ-avignon.fr – English coming soon
  - Speech group (~10 permanent and 20 PhD students)
    - Dialogue systems (Renato De Mori)
    - Speaker id/diarization (Alize toolkit, Jean-François Bonastre)
    - STT: French and resource-sparse languages
    - Voice/Language pathologies

Outline

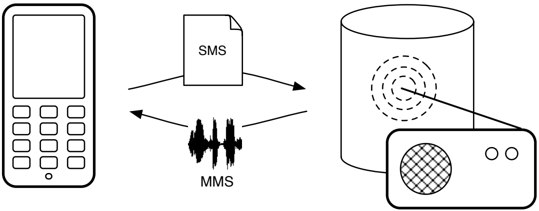
1. Introduction
2. Speech Database Browsing
   - Context
   - Interactive timeline
3. Prototype
   - Demo
   - Implementation
   - Performance
4. Conclusion

Outline

1. Introduction
2. Speech Database Browsing
   - Context
   - Interactive timeline
3. Prototype
   - Demo
   - Implementation
   - Performance
4. Conclusion

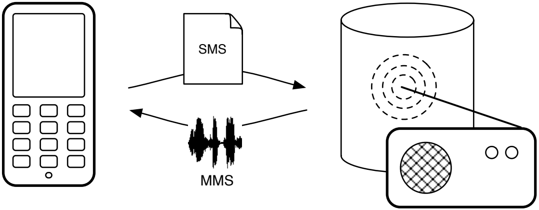
Application context: spoken information retrieval

- SMS: text based query (eg. "baseball results")
- Generate a **spoken summary** of the news
- Audio delivered by MMS



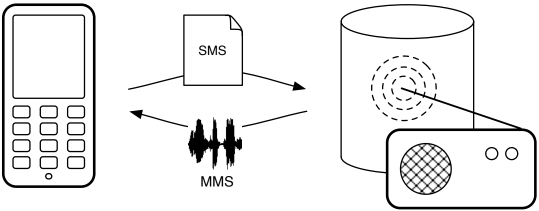
Application context: spoken information retrieval

- SMS: text based query (eg. "baseball results")
- Generate a **spoken summary** of the news
- Audio delivered by MMS



Application context: spoken information retrieval

- SMS: text based query (eg. "baseball results")
- Generate a **spoken summary** of the news
- Audio delivered by MMS



Approaches

- Knowledge rich
  - Database of information items
  - Text generation
  - Speech synthesis
- Open domain (data driven)
  - Collect broadcast news (or/and other sources)
  - Select informative segments (sentences)
  - Segment playback
- Hybrid
  - Fill the knowledge base from collected BN
  - Contextualize the segment playback with speech synthesis
  - ...

Approaches

- Knowledge rich
  - Database of information items
  - Text generation
  - Speech synthesis
- Open domain (data driven)
  - Collect broadcast news (or/and other sources)
  - Select informative segments (sentences)
  - Segment playback
- Hybrid
  - Fill the knowledge base from collected BN
  - Contextualize the segment playback with speech synthesis
  - ...

Approaches

- Knowledge rich
  - Database of information items
  - Text generation
  - Speech synthesis
- Open domain (data driven)
  - Collect broadcast news (or/and other sources)
  - Select informative segments (sentences)
  - Segment playback
- Hybrid
  - Fill the knowledge base from collected BN
  - Contextualize the segment playback with speech synthesis
  - ...

Approaches

- Knowledge rich
  - Database of information items
  - Text generation
  - Speech synthesis
- Open domain (data driven)
  - Collect broadcast news (or/and other sources)
  - Select informative segments (sentences)
  - Segment playback
- Hybrid
  - Fill the knowledge base from collected BN
  - Contextualize the segment playback with speech synthesis
  - ...

From text to speech summarization

- Rich transcription
  - Acoustic segmentation, diarization
  - Speech-to-text transcript
  - Information extraction
- Summarization by sentence selection
  - Impact of STT errors (and other RT errors)
  - Require accurate sentence boundaries
  - Perception of “cut-and-paste”
- Audio only features
  - Speaker state and identity
  - Emphasis
  - Speech quality

From text to speech summarization

- Rich transcription
  - Acoustic segmentation, diarization
  - Speech-to-text transcript
  - Information extraction
- Summarization by sentence selection
  - Impact of STT errors (and other RT errors)
  - Require accurate sentence boundaries
  - Perception of “cut-and-paste”
- Audio only features
  - Speaker state and identity
  - Emphasis
  - Speech quality

From text to speech summarization

- Rich transcription
  - Acoustic segmentation, diarization
  - Speech-to-text transcript
  - Information extraction
- Summarization by sentence selection
  - Impact of STT errors (and other RT errors)
  - Require accurate sentence boundaries
  - Perception of "cut-and-paste"
- Audio only features
  - Speaker state and identity
  - Emphasis
  - Speech quality

My work at LIA

- Setup a rich transcription processing chain
  - Speeral toolkit for STT
  - Alize platform for diarization
  - Word lattice based NE tagging
  - CRF based Sentence Segmentation
- Build and evaluate a text summarization system
  - MMR-LSA summarization system
  - Document Understanding Conference (DUC) evaluation
  - Impact on audio: simulate ASR
- Study possible user interactions
  - Speech database browsing
  - Interactive timeline
- Next PhD student: Audio only features

My work at LIA

- Setup a rich transcription processing chain
  - Speeral toolkit for STT
  - Alize platform for diarization
  - Word lattice based NE tagging
  - CRF based Sentence Segmentation
- Build and evaluate a text summarization system
  - MMR-LSA summarization system
  - Document Understanding Conference (DUC) evaluation
  - Impact on audio: simulate ASR
- Study possible user interactions
  - Speech database browsing
  - Interactive timeline
- Next PhD student: Audio only features

My work at LIA

- Setup a rich transcription processing chain
  - Speeral toolkit for STT
  - Alize platform for diarization
  - Word lattice based NE tagging
  - CRF based Sentence Segmentation
- Build and evaluate a text summarization system
  - MMR-LSA summarization system
  - Document Understanding Conference (DUC) evaluation
  - Impact on audio: simulate ASR
- Study possible user interactions
  - Speech database browsing
  - Interactive timeline
- Next PhD student: Audio only features

My work at LIA

- Setup a rich transcription processing chain
  - Speeral toolkit for STT
  - Alize platform for diarization
  - Word lattice based NE tagging
  - CRF based Sentence Segmentation
- Build and evaluate a text summarization system
  - MMR-LSA summarization system
  - Document Understanding Conference (DUC) evaluation
  - Impact on audio: simulate ASR
- Study possible user interactions
  - Speech database browsing
  - Interactive timeline
- Next PhD student: Audio only features

Outline

Constraints

- Continuous audio archives (BN, Meetings...)
  - "Decades" of recordings
  - Multiple sources
- Why isn't "raw" summarization suitable?
  - Reintroduce context
  - Track the source
- Information retrieval → exploration
  - Structure discovery
  - Temporal vs Topical structure
- Speech is bound to time
  - Wait to hear more
  - No static representation

Constraints

- Continuous audio archives (BN, Meetings...)
  - "Decades" of recordings
  - Multiple sources
- Why isn't "raw" summarization suitable?
  - Reintroduce context
  - Track the source
- Information retrieval $\longrightarrow$ exploration
  - Structure discovery
  - Temporal vs Topical structure
- Speech is bound to time
  - Wait to hear more
  - No static representation

Constraints

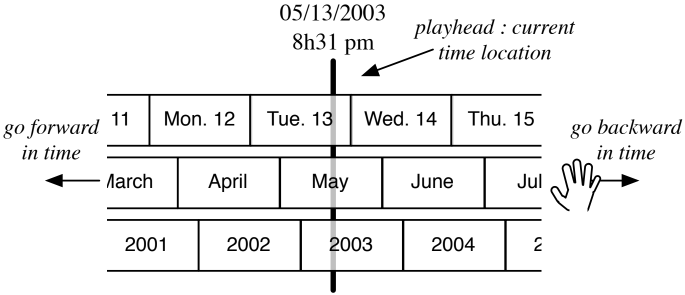
- Continuous audio archives (BN, Meetings...)
  - "Decades" of recordings
  - Multiple sources
- Why isn't "raw" summarization suitable?
  - Reintroduce context
  - Track the source
- Information retrieval $\rightarrow$ exploration
  - Structure discovery
  - Temporal vs Topical structure
- Speech is bound to time
  - Wait to hear more
  - No static representation

Constraints

- Continuous audio archives (BN, Meetings...)
  - "Decades" of recordings
  - Multiple sources
- Why isn't "raw" summarization suitable?
  - Reintroduce context
  - Track the source
- Information retrieval → exploration
  - Structure discovery
  - Temporal vs Topical structure
- Speech is bound to time
  - Wait to hear more
  - No static representation

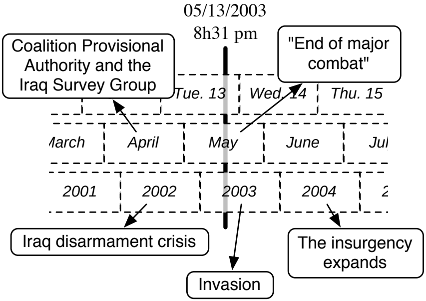
Multiscale playhead

- Synchronous multiscale timeline
  - Slices representing years, months, days...
  - Dragging one slice synchronize the others
  - Easy "time travel" at every granularity
- Annotation



Multiscale playhead

- Synchronous multiscale timeline
- Annotation
  - Need for structure information
  - Topic/Event labels
  - Example from Wikipedia (Iraq war)

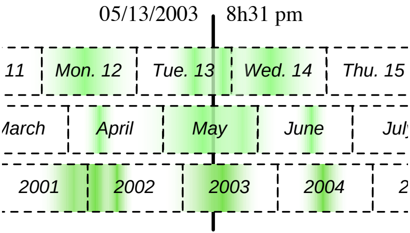


Automatic Annotation

- Constraints
  - Reflect a user query
  - Highlight regions of interest
  - Interactive
- Approach
  - Relevance density (information retrieval)
  - Anchorage points (automatic summarization)

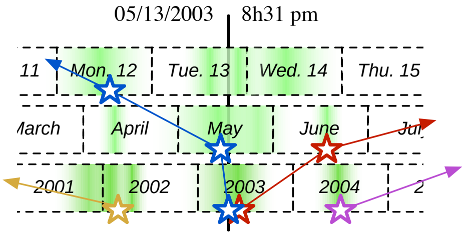
Automatic Annotation

- Constraints
  - Reflect a user query
  - Highlight regions of interest
  - Interactive
- Approach
  - Relevance density (information retrieval)
  - Anchorage points (automatic summarization)



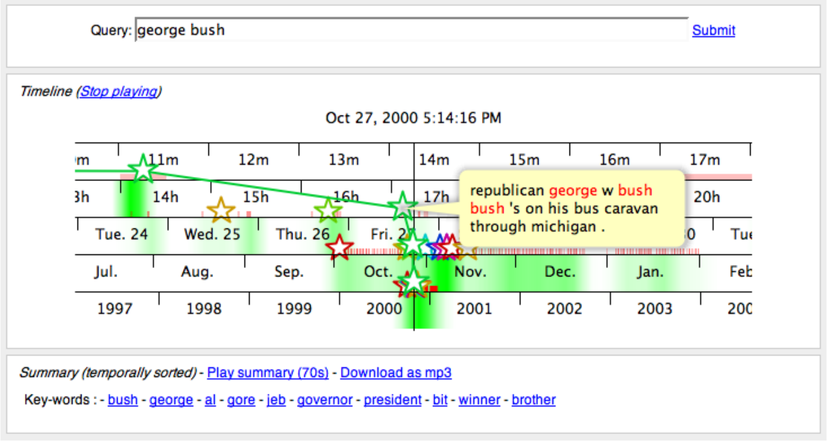
Automatic Annotation

- Constraints
  - Reflect a user query
  - Highlight regions of interest
  - Interactive
- Approach
  - Relevance density (information retrieval)
  - Anchorage points (automatic summarization)



Outline

Screen capture (and demo if lucky)

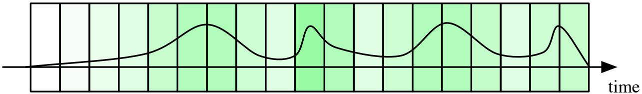


Information density

- *n* highest-relevant sentences
- Okapi IR model *[Robertson et al]*,

$$\frac{P(R|D,Q)}{P(\overline{R}|D,Q)} \sim \prod_{w} \frac{P_w(1-\overline{P_w})}{\overline{P_w}(1-P_w)} \sim \sum_{w} \log f(w, D, \Lambda)$$

- Stop-word removal
- Context modeling (interpolation with neighboring sentences)

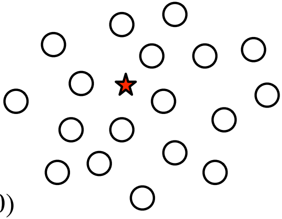


Anchorage points: Maximal Marginal Relevance (MMR)

- Select the *m* highest-representative sentences
- Greedy sentence selection *[Goldstein et al]*

$$(\hat{\mathbf{s}})_{k+1} = \underset{\mathbf{s_i} \notin mmr_k}{\text{argmax}} \left(\lambda coverage(\mathbf{s_i}, \mathbf{q}) - (1-\lambda) \max_{\mathbf{s_j} \in mmr_k} redundacy(\mathbf{s_i}, \mathbf{s_j}) \right)$$

- Duration based stopping criterion



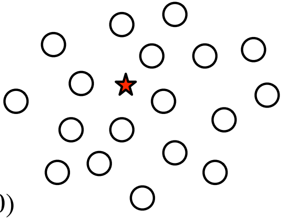
(0)

Anchorage points: Maximal Marginal Relevance (MMR)

- Select the *m* highest-representative sentences
- Greedy sentence selection *[Goldstein et al]*

$$(\hat{\mathbf{s}})_{k+1} = \underset{\mathbf{s_i} \notin mmr_k}{\operatorname{argmax}} \left(\lambda coverage(\mathbf{s_i}, \mathbf{q}) - (1-\lambda) \max_{\mathbf{s_j} \in mmr_k} redundacy(\mathbf{s_i}, \mathbf{s_j}) \right)$$

- Duration based stopping criterion



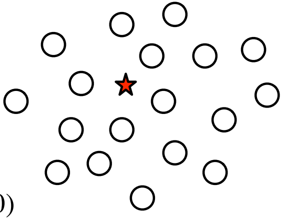
(0)

Anchorage points: Maximal Marginal Relevance (MMR)

- Select the *m* highest-representative sentences
- Greedy sentence selection *[Goldstein et al]*

$$(\hat{\mathbf{s}})_{k+1} = \underset{\mathbf{s_i} \notin mmr_k}{\mathrm{argmax}} \left(\lambda coverage(\mathbf{s_i}, \mathbf{q}) - (1-\lambda) \max_{\mathbf{s_j} \in mmr_k} redundacy(\mathbf{s_i}, \mathbf{s_j}) \right)$$

- Duration based stopping criterion



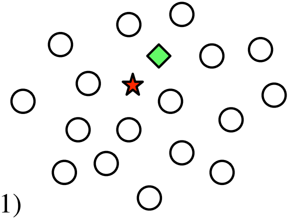
(0)

Anchorage points: Maximal Marginal Relevance (MMR)

- Select the *m* highest-representative sentences
- Greedy sentence selection *[Goldstein et al]*

$$(\hat{\mathbf{s}})_{k+1} = \underset{\mathbf{s_i} \notin mmr_k}{\text{argmax}} \left(\lambda coverage(\mathbf{s_i}, \mathbf{q}) - (1-\lambda) \max_{\mathbf{s_j} \in mmr_k} redundacy(\mathbf{s_i}, \mathbf{s_j}) \right)$$

- Duration based stopping criterion



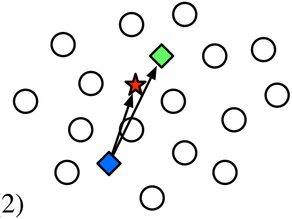
(1)

Anchorage points: Maximal Marginal Relevance (MMR)

- Select the *m* highest-representative sentences
- Greedy sentence selection *[Goldstein et al]*

$$(\hat{\mathbf{s}})_{k+1} = \underset{\mathbf{s_i} \notin mmr_k}{\operatorname{argmax}} \left(\lambda coverage(\mathbf{s_i}, \mathbf{q}) - (1-\lambda) \max_{\mathbf{s_j} \in mmr_k} redundacy(\mathbf{s_i}, \mathbf{s_j}) \right)$$

- Duration based stopping criterion



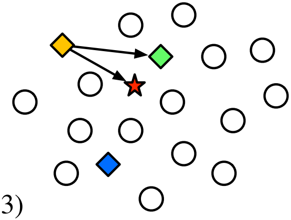
(2)

Anchorage points: Maximal Marginal Relevance (MMR)

- Select the *m* highest-representative sentences
- Greedy sentence selection *[Goldstein et al]*

$$(\hat{\mathbf{s}})_{k+1} = \underset{\mathbf{s_i} \notin mmr_k}{\text{argmax}} \left(\lambda coverage(\mathbf{s_i}, \mathbf{q}) - (1-\lambda) \max_{\mathbf{s_j} \in mmr_k} redundacy(\mathbf{s_i}, \mathbf{s_j}) \right)$$

- Duration based stopping criterion



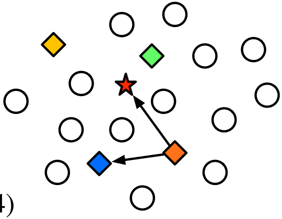
(3)

Anchorage points: Maximal Marginal Relevance (MMR)

- Select the *m* highest-representative sentences
- Greedy sentence selection *[Goldstein et al]*

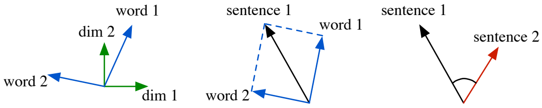
$$(\hat{\mathbf{s}})_{k+1} = \underset{\mathbf{s_i} \notin mmr_k}{\operatorname{argmax}} \left(\lambda coverage(\mathbf{s_i}, \mathbf{q}) - (1-\lambda) \max_{\mathbf{s_j} \in mmr_k} redundacy(\mathbf{s_i}, \mathbf{s_j}) \right)$$

- Duration based stopping criterion


(4)

Latent Semantic Analysis (LSA)

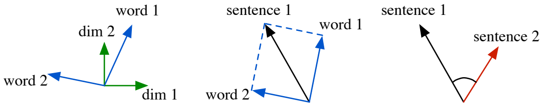
- Similarity between sentences (Generalized VSM)
  "Chris purchased a BMW"
  "Mr. Jones bought a car"



- Cooccurrence matrix (lexicon $\times$ lexicon, sliding window)
  - Train on a big corpus *[Peters et al]*
  - Reduce the matrix by SVD, $X^* = U\Sigma_k V^T$
  - Project sentences, $\mathbf{s}^* = \Sigma_k^{-1} U^T \mathbf{s}$
  - Cosine similarity, $cosine(a, b) = \frac{a \cdot b}{|a||b|}$

Latent Semantic Analysis (LSA)

- Similarity between sentences (Generalized VSM)
  "Chris purchased a BMW"
  "Mr. Jones bought a car"



- Cooccurrence matrix (lexicon $\times$ lexicon, sliding window)
  - Train on a big corpus *[Peters et al]*
  - Reduce the matrix by SVD, $X^* = U\Sigma_k V^T$
  - Project sentences, $\mathbf{s}^* = \Sigma_k^{-1} U^T \mathbf{s}$
  - Cosine similarity, $cosine(a, b) = \frac{a \cdot b}{|a||b|}$

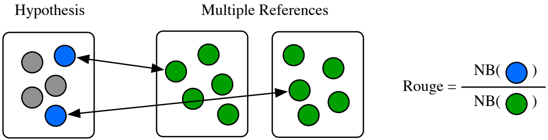
Performance

- ESTER 2005 Evaluation (French BN)

Task	Perf.	Measure
Speech detection	99	F_1-m
Speech+Music det.	92	F_1-m
Music detection	54	F_1-m
Diarization	19	%err
STT	22	WER
Sentence Segmentation	68	F_1-m
Named Entities	63	F_1-m

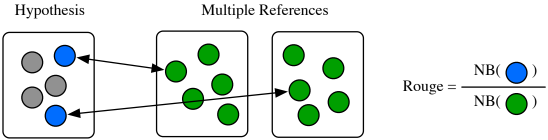
Document Understanding Evaluation

- Multidocument, user oriented, text summarization
  - 50 topics, 25 newswire documents per topic
  - Human judgments (linguistic quality and responsiveness)
  - Automatic judgments (not a trivial at all)
- ROUGE
  - Recall in *n*-grams with a set of hand written summaries
  - Correlated with Human judgements



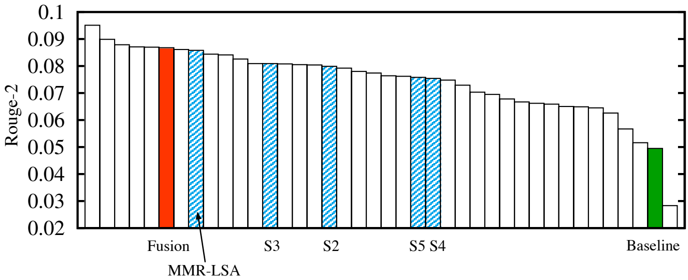
Document Understanding Evaluation

- Multidocument, user oriented, text summarization
  - 50 topics, 25 newswire documents per topic
  - Human judgments (linguistic quality and responsiveness)
  - Automatic judgments (not a trivial at all)
- ROUGE
  - Recall in *n*-grams with a set of hand written summaries
  - Correlated with Human judgements



DUC Results on text documents

- LIA submission at DUC 2006, 2007
  - Fusion of up to 7 (sentence ranking) systems
  - A lot of heuristics, linguistic pre/post processing



Simulating a spoken content

- Simulated STT on DUC documents
  - Uniform random errors
  - Worst case for a summarizer
- Conditions
  - Noisy: word errors appear in the summary
  - Cleaned: only sentence selection is affected

Degradation	WER	R2 Noisy		R2 Cleaned	
None	0.0	0.08407		0.08407	
Replace OOV	1.0	0.08255	-1.8%	0.08318	-1.0%
Remove OOV	1.0	0.08283	-1.4%	0.08279	-1.5%
Replace NE	10.4	0.06741	-19.8%	0.08029	-4.4%
Remove NE	10.4	0.07211	-14.2%	0.07991	-4.9%
Random errors	10.0	0.07440	-11.5%	0.08232	-2.0%

Simulating a spoken content

- Simulated STT on DUC documents
  - Uniform random errors
  - Worst case for a summarizer
- Conditions
  - Noisy: word errors appear in the summary
  - Cleaned: only sentence selection is affected

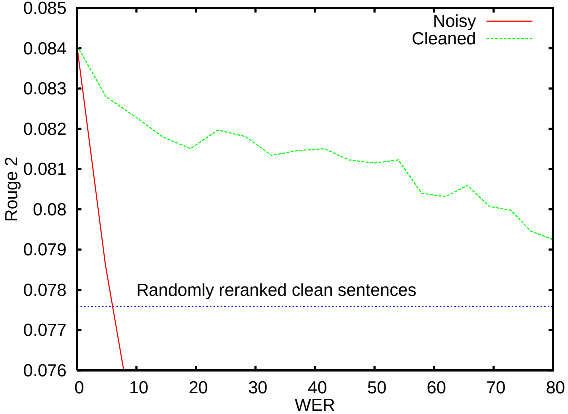
Degradation	WER	R2 Noisy		R2 Cleaned	
None	0.0	0.08407		0.08407	
Replace OOV	1.0	0.08255	-1.8%	0.08318	-1.0%
Remove OOV	1.0	0.08283	-1.4%	0.08279	-1.5%
Replace NE	10.4	0.06741	-19.8%	0.08029	-4.4%
Remove NE	10.4	0.07211	-14.2%	0.07991	-4.9%
Random errors	10.0	0.07440	-11.5%	0.08232	-2.0%

Simulating a spoken content

- Simulated STT on DUC documents
  - Uniform random errors
  - Worst case for a summarizer
- Conditions
  - Noisy: word errors appear in the summary
  - Cleaned: only sentence selection is affected

Degradation	**WER**	**R2 Noisy**		**R2 Cleaned**	
None	0.0	0.08407		0.08407	
Replace OOV	1.0	0.08255	-1.8%	0.08318	-1.0%
Remove OOV	1.0	0.08283	-1.4%	0.08279	-1.5%
Replace NE	10.4	0.06741	-19.8%	0.08029	-4.4%
Remove NE	10.4	0.07211	-14.2%	0.07991	-4.9%
Random errors	10.0	0.07440	-11.5%	0.08232	-2.0%

Rouge-2 / WER



Head-Baseline: $Rouge2 = 0.049$

Random-Baseline: $Rouge2 = 0.055$

Outline

Conclusion and future work

- Improving speech database browsing
  - Multi-scale interactive timeline
  - Annotation using IR and Automatic Summarization techniques
- Future work
  - Evaluation (ergonomics and relevance)
  - Topical dimension: representation, exploration
  - Label formulation
  - Timeline of discourse $\longrightarrow$ Timeline of events
  - Indirect/Passive querying

Conclusion and future work

- Improving speech database browsing
  - Multi-scale interactive timeline
  - Annotation using IR and Automatic Summarization techniques
- Future work
  - Evaluation (ergonomics and relevance)
  - Topical dimension: representation, exploration
  - Label formulation
  - Timeline of discourse $\rightarrow$ Timeline of events
  - Indirect/Passive querying