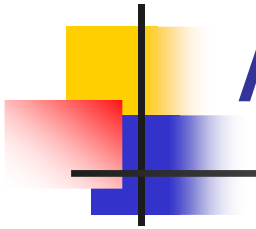
Information Retrieval from Automatic Speech Transcripts

Diana Inkpen
University of Ottawa, SITE

Browsing spoken audio data

- Ways to facilitate it:
  - gist a spoken audio document by glancing over a transcript generated through Automatic Speech Recognition (ASR).
  - look at keyphrases extracted from these transcripts.
  - text information retrieval from ASR transcripts;

- Unfortunately, the transcripts typically contain many recognition errors which are highly distracting and make gisting more difficult.



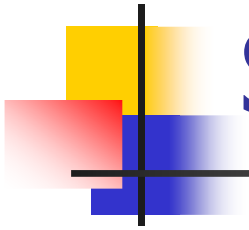
ARISE Project

- "Memories of Synchronicity: Knowledge Management and Visualization of Interaction Transcripts in Innovative Collaboration Environments"
  - Universities of Toronto, Waterloo, Ottawa, NRC, IBM
  - Tools for effective collaboration
  - Recordings of meetings, lectures, etc.
  - Use speech recognition to facilitate navigation through the material:
    - information retrieval on automatic transcripts
    - keyphrase summaries for browsing

Previous work

- Hirschberg et al. (1999), and Nakatani et al. (1998): use automatic transcripts for gisting and navigating audio documents.
- Text-based summarization techniques on automatic speech transcripts.
  - For example, keyphrases extracted from automatic transcripts (Désilets *et al.* 2001).
- Semantic similarity measures were used for various tasks (Budanitsky & Hirst 2001) (Jarmasz & Szpakowicz 2003) (Pedersen *et al.* 2004).

- Information retrieval: TREC SDR, CLEF CL-SR

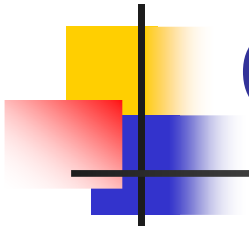


Speech to keyphrases

- Désilets *et al.* (2001) used speech recognition to produce automatic transcripts, then extracted keyphrases with Extractor (Turney 2000).
- Accurate keyphrases for transcriptions with Word Error Rates (WER) of 25%
- Performance was less than ideal for transcripts with WER of 60%.

Semantic outliers in keyphrases

- A **keyphrase** consists of one, two, or three **keywords**. Examples: *Russian cities, river, elated, nazis, war, scene, stanza*
- The keywords word error rate (cWER) is much lower than WER in speech transcripts.
- The transcription errors that are in the keyphrases are semantically unrelated to the other words in keyphrases. Low semantic coherence with neighbors.

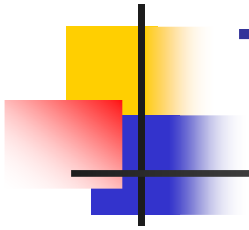


Goals

1. Filter out (replace with placeholders) semantic outliers from automatic speech transcripts.

2. Filter out (remove) semantic outliers to improve the quality of keyphrases.

3. Information retrieval system on transcripts.



The data (goals 1 and 2)

- 100 stories from the TDT2 English Audio data
- Correct transcripts generated by humans.
- Two types of automatically-generated speech transcripts (two datasets):
  - NIST/BBN time-adaptive speech recognizer: moderate WER 27.6% – broadcast quality
  - Dragon NaturallySpeaking speaker dependant recognizer (not trained): high WER 62.3%

    – simulate less than broadcast quality

Manual transcript: Time now for our geography quiz today. We're traveling down the Volga river to a city that, like many Russian cities, has had several names. But this one stands out as the scene of an epic battle in world war two in which the Nazis were annihilated.

Keyphrases: *Russian cities, city, Volga river, Nazis, war, epic battle, scene*

BBN transcript: time now for a geography was they were traveling down river to a city that like many russian cities has had several names but this one stanza is the scene of ethnic and national and world war two in which the nazis were nine elated

Keyphrases: *Russian cities, city, river, elated, nazis, war, scene, stanza*

Detected outlier keywords: ***stanza, elated***

Detecting outliers in the speech transcripts

Manual transcript: Time now for our geography quiz today. We're traveling down the Volga river to a city that, like many Russian cities, has had several names. But this one stands out as the scene of an epic battle in world war two in which the Nazis were annihilated.

BBN transcript, without the semantic outliers: time now for a geography was they were traveling down river to a city that like many russian cities has had several names but this one …… is the scene of ethnic and national and world war two in which the nazis were nine …….

Detected outliers: ***stanza, elated***

Filtering semantic outliers in speech transcripts

Original spoken text: "We need to decide quickly whether we will go for a large expensive plasma screen or for a bunch of smaller and cheaper ones and tile them together."

T1: "Weenie to decide quickly whether local for large expensive plasma screen aura for a bunch of smaller and cheaper ones and Holland together"

T2: "... ... decide quickly whether ... large expensive plasma screen ... for a bunch of smaller and cheaper ones and ... together"

- T2 would allow the user to more quickly and more accurately get the gist of what was said

Semantic outliers in transcripts

- **Use the set of content words in the automatic transcript**
- All content words in the document (or segment) vs. Context window of 20 words
- Try various semantic similarity measures (corpus-based, thesaurus-based)
- Use the closest neighbour, the 3-closest neighbours, or all of them.

- For semantic outliers in keyphrases, use the same method, with **the set of words in the keyphrases**.

The method - For each content word w in the automatic transcript

1. Compute the **neighborhood** N(w), i.e. the set of content words that occur "close" to w in the transcript (include w).

2. Compute **pair-wise semantic similarity** scores $S(w_i,w_j)$ between all pairs of words $w_i \neq w_j$ in N(w), using a semantic similarity measure.

3. Compute the **semantic coherence** $SC(w_i)$ by "aggregating" the pair-wise semantic similarities $S(w_i, w_j)$ of w_i with all its neighbors $w_j \neq w_i$ in N(w).

4. Let SC_{avg} be the average of $SC(w_i)$ over all w_i in the neighborhood N(w).

5. Label w as a recognition errors if $SC(w) < K\ SC_{avg}$.

Semantic outliers in transcripts Variants

- **Step1Variant-Step2Variant-Step3Variant**.
  - Step1 – N(w) = All words vs. Window of 20 words
  - Step2 – PMI vs. Roget similarity measure
  - Step3 – AVG vs. MAX vs. 3MAX
- 2 x 2 x 3 = 12 possible combinations.
- For example, All-PMI-AVG means the configuration that uses All in Step 1, PMI in Step 2, and AVG in Step 3.

Semantic similarity

	Dictionary-based		**Corpus-based**		
	Leacock & Chodorow (WN)	**Roget** (Roget)	**Cosine** (BNC)	**Correla-tion** (BNC)	**PMI** (MultiT)
M&C	0.821	0.878	0.406	0.438	0.759
R&G	0.852	0.818	0.472	0.517	0.746

=> We used PMI and Roget-based similarity

Semantic similarity - PMI

- The semantic similarity score between two words w1 and w2 is their pointwise mutual information score.

- $S(w1, w2) = \log P(w1, w2) / P(w1)\, P(w2)$
  $= \log C(w1, w2)\, N / C(w1)\, C(w2)$

- The corpus = 1 terabyte of Web data; the Waterloo Multitext system (Clarke and Terra 2003).

Evaluation of filtered transcripts

Recognition error detection as a classification task

	Correctly transcribed (actual)	NOT Correctly transcribed (actual)
Correctly transcribed (predicted)	True Positive	False Positive (remaining WER)
NOT Correctly transcribed (predicted)	False Negative (lost)	True Negative (semantic outliers)

Evaluation measures

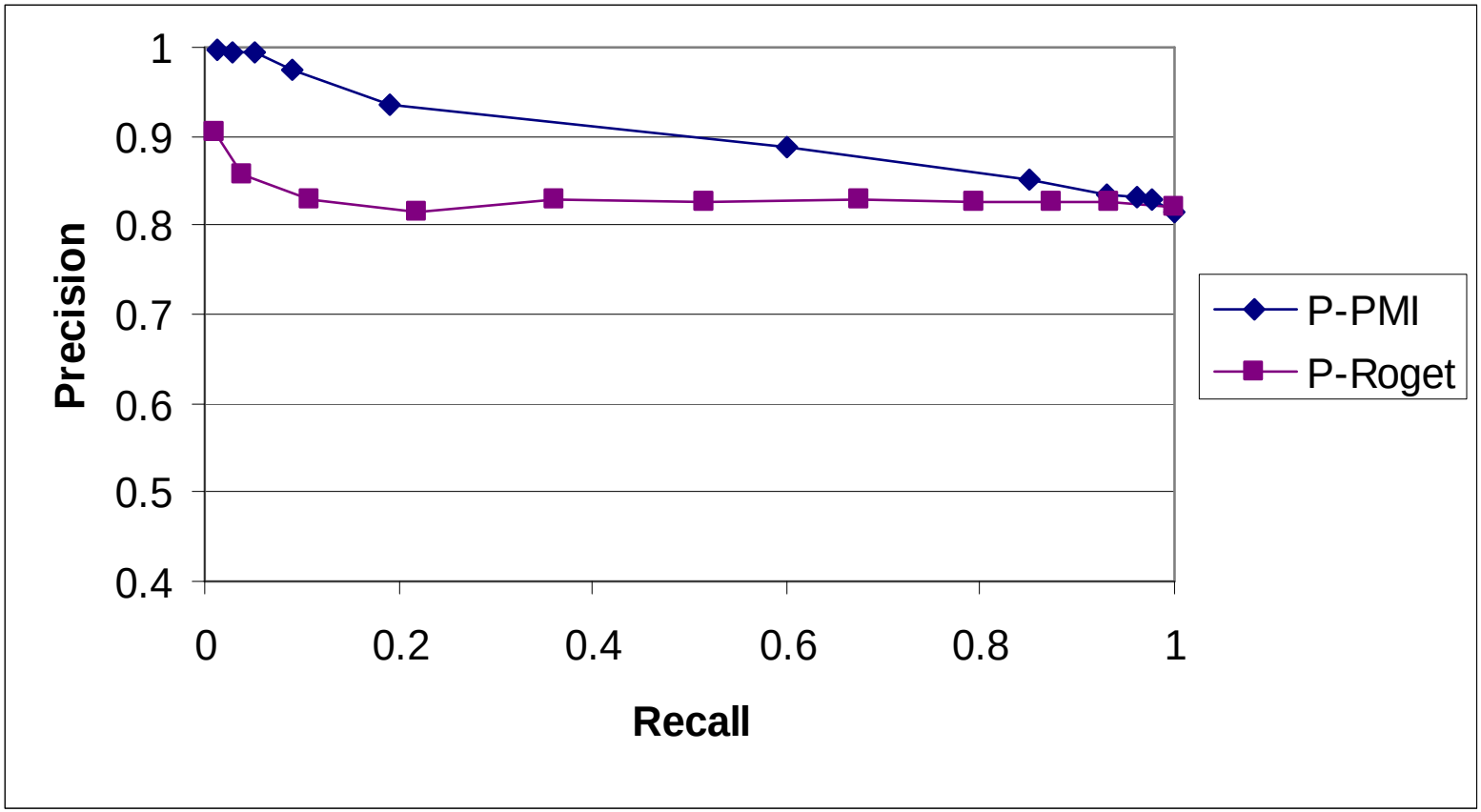
- WER = word error rate (deletions, insertions, substitutions)
- cWER = word error rate for content words only (no penalty for deletions)

  = % of content words that are in the automatic transcript but not in the manual transcript

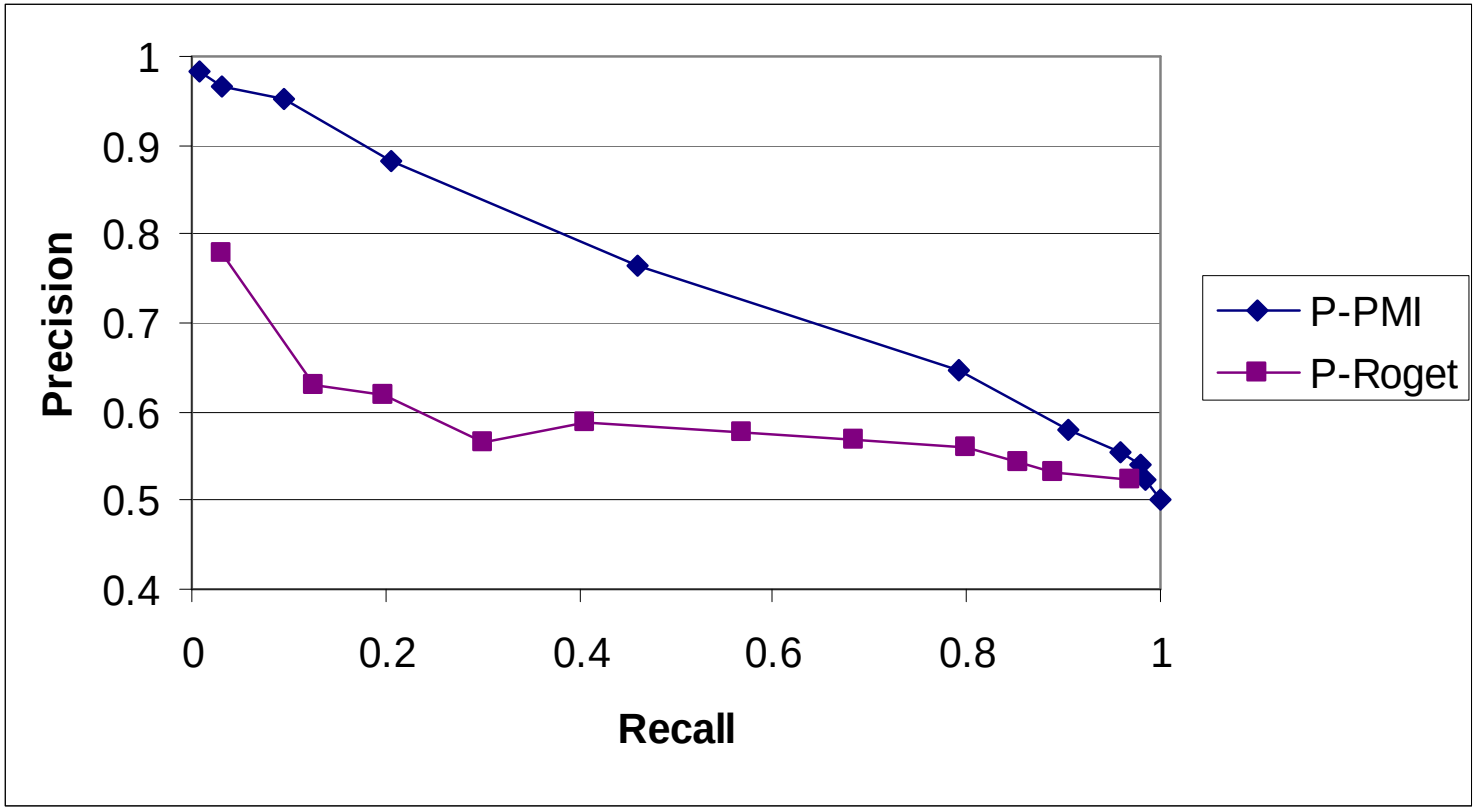
  = FP / (TP + FP)
- %Lost = % of content words that are eliminated but they are in the manual transcript

  = FN / (TP + FN)
- Precision = TP / TP + FP, Recall = TP / TP + FN,
- F measure = 2PR/(P+R)

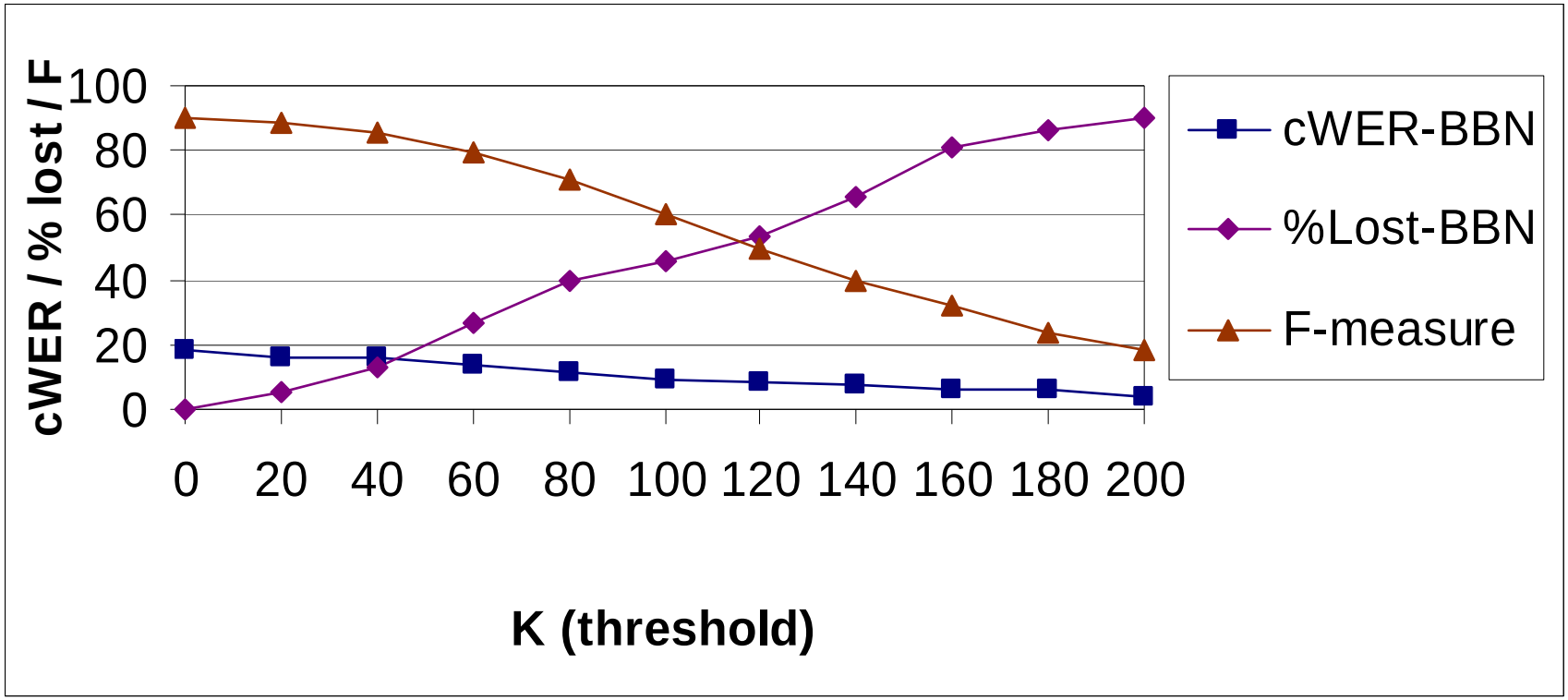
P-R curves of PMI vs. Roget (with All and AVG) on the BBN dataset. Each P-R point corresponds to a different value of the threshold K (high Recall for low values of K, high Precision for high values of K).



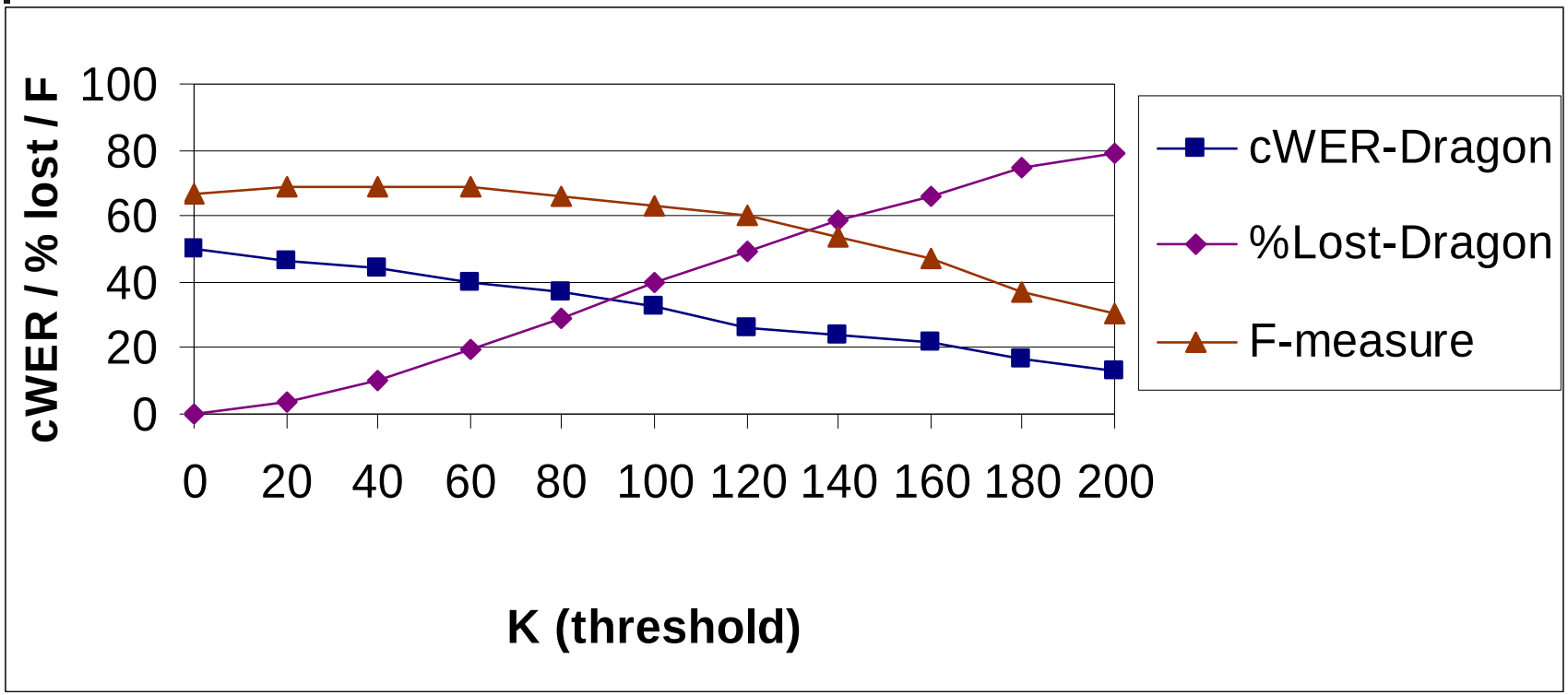
P-R curves of PMI vs. Roget (with All and AVG) on the Dragon dataset



Content Word Error Rate (cWER), %Lost good words (%Lost) and F-measure as a function of the filtering level K for the Window-PMI-3MAX configuration on the BBN dataset

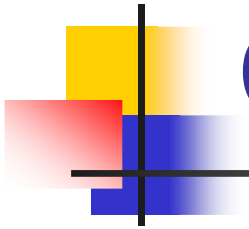


Content Word Error Rate (cWER), %Lost good words (%Lost) and F-measure as a function of the filtering level K for the Window-PMI-3MAX configuration on the Dragon dataset.



Best Variant: Window-PMI-3MAX

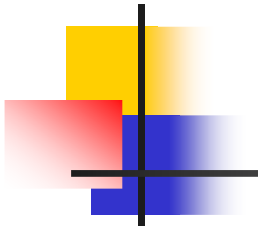
Tran-scripts	WER transcripts	cWER initially	Filtered transcripts K	%Lost	cWER	Reduction
BBN	27.6%	18.3%	100	45%	9.15%	50%
Dragon	62.3%	49.3%	120	50%	25.65%	50%



Comparison to previous work

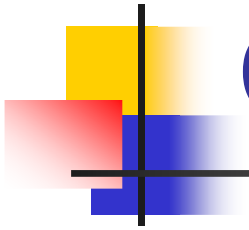
Cox and Dasmahapatra (2000) LSA-based measure

- Our PMI-based measure seems to performs better.
- At P=90%, they obtain R=12%, we obtain R=20%.
- At P=80%, they obtain R=50%, we get R=100%.
- Not completely comparable: different audio corpora (WSJCAM0 vs. TDT2), similar initial WERs (30%).
- LSA measure was computed based on a corpus similar to the audio corpus used for evaluation (WSJ). Our PMI measure was computed on a general sample of WWW, not tailored to the audio corpus used for evaluation.



Information Retrieval for a Spontaneous Conversational Speech Collection (goal 3)

- Cross-Language Evaluation Forum (CLEF) 2005, 2006
  - Cross-Language Speech Retrieval (CL-SR) track



CLEF 2005: CL-SR Task

Collection – oral testimonies collected by the Shoah Foundation Institute for Visual History and Education

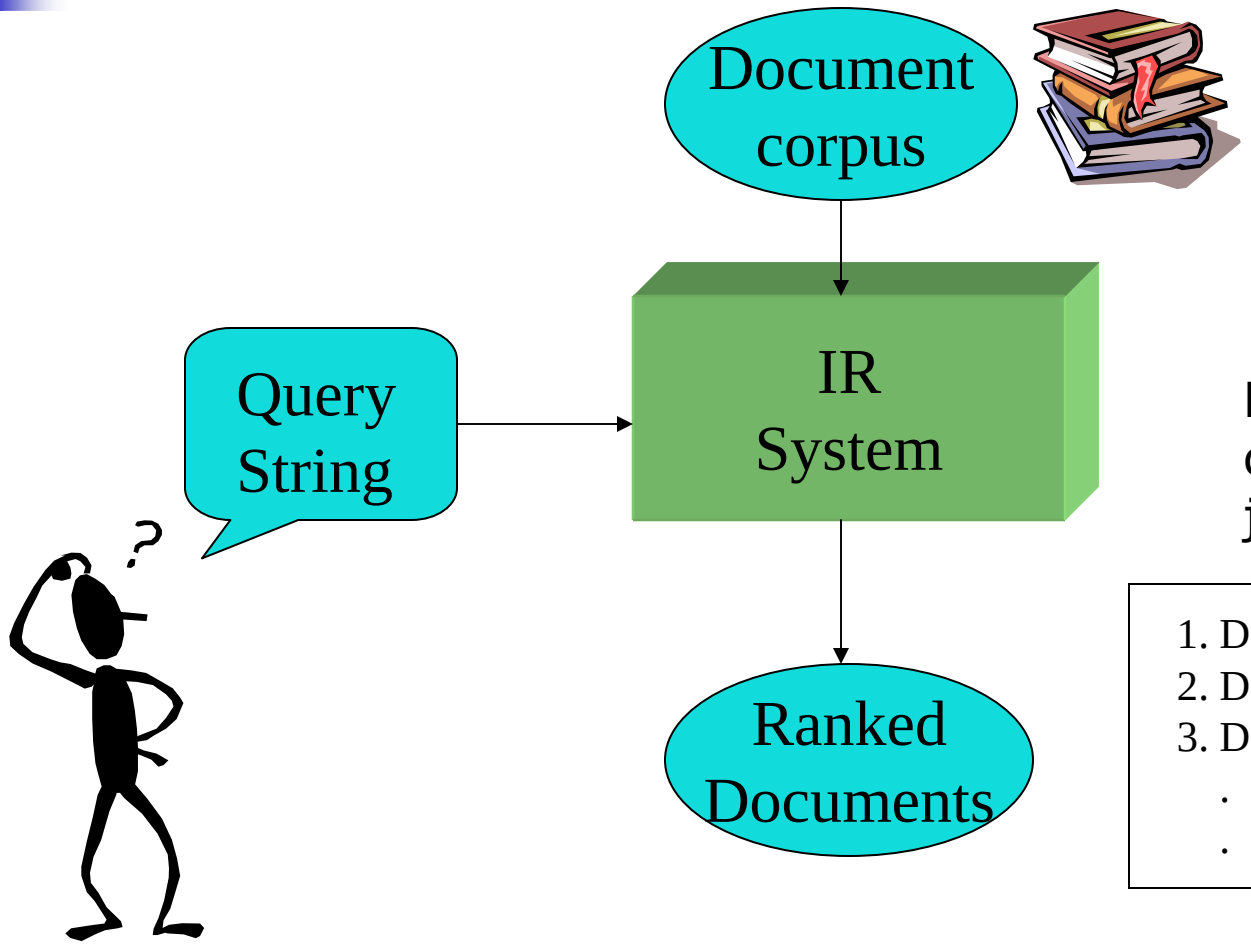
- ASR transcribed text (WER 38%)

  8,104 segments, from 272 interviews with Holocaust survivors, totaling 589 hours of speech

- automatic keywords, manual keywords and 3-line summaries

- Training queries (38), test queries (25) – actual user requests
- Relevance judgments

IR System Architecture

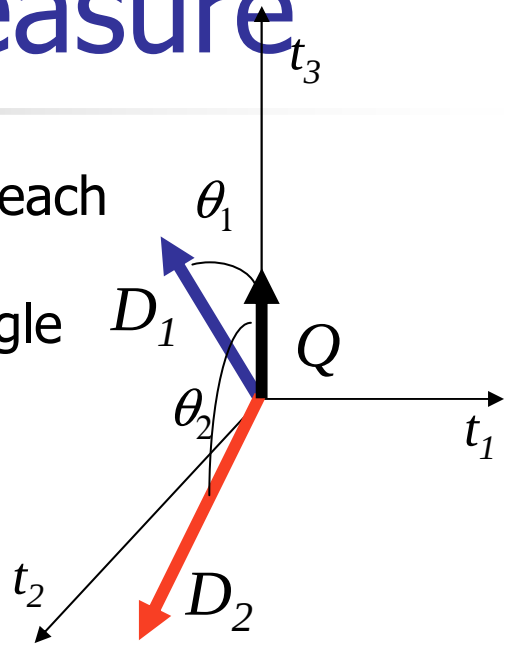


For evaluation, compare to relevance judgements

Vector Space Model: Cosine Similarity Measure

- Measure the similarity between the query and each document.
- Cosine similarity measures the cosine of the angle between two vectors.

$$\text{CosSim}(\boldsymbol{d}_j, \boldsymbol{q}) = \frac{\vec{d}_j \cdot \vec{q}}{\left|\vec{d}_j\right| \cdot \left|\vec{q}\right|} = \frac{\sum_{i=1}^{t} (w_{ij} \cdot w_{iq})}{\sqrt{\sum_{i=1}^{t} w_{ij}^2 \cdot \sum_{i=1}^{t} w_{iq}^2}}$$



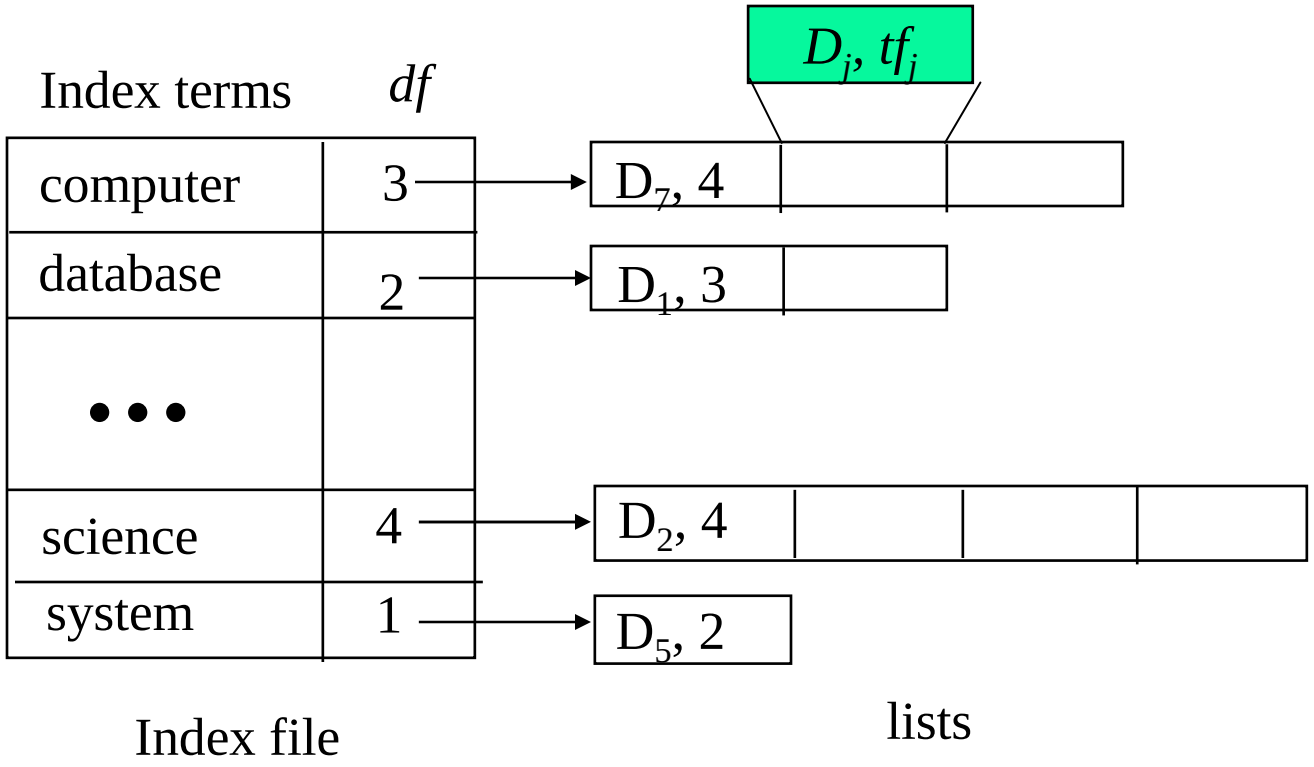
$D_1 = 2T_1 + 3T_2 + 5T_3$ $\quad \text{CosSim}(D_1, Q) = 10 / \sqrt{(4+9+25)(0+0+4)} = 0.81$

$D_2 = 3T_1 + 7T_2 + 1T_3$ $\quad \text{CosSim}(D_2, Q) = 2 / \sqrt{(9+49+1)(0+0+4)} = 0.13$

$Q = 0T_1 + 0T_2 + 2T_3$

D_1 is 6 times better than D_2 using cosine similarity.

Indexing (tf-idf)



Evaluation: Mean average precision
(example of computing Recall/Precision points)

n	doc #	relevant
1	588	x
2	589	x
3	576	
4	590	x
5	986	
6	592	x
7	984	
8	988	
9	578	
10	985	
11	103	
12	591	
13	772	x
14	990	

Let total # of relevant docs = 6
Check each new recall point:

R=1/6=0.167; P=1/1=1

R=2/6=0.333; P=2/2=1

R=3/6=0.5; P=3/4=0.75

R=4/6=0.667; P=4/6=0.667

R=5/6=0.833; P=5/13=0.38

Missing one relevant document. Never reach 100% recall

Our system

- **SMART IR system** (Buckley et al, 1993)
- **Terrier** (Amati and van Rijsbergen 2002) (Ounis et al 2005)

- For cross-language system, use online MT tools

Spanish, German, French:

1. http://www.google.com/language_tools?hl=en
2. http://www.babelfish.altavista.com
3. http://freetranslation.com
4. http://www.wordlingo.com/en/products_services/ wordlingo_translator.html
5. http://www.systranet.com/systran/net
6. http://www.online-translator.com/srvurl.asp?lang=en
7. http://www.freetranslation.paralink.com

Czech:

1. http://intertran.tranexp.com/Translate/result.shtml

Example document (Segment)

<DOC>
<DOCNO>VHF00195-073439.026</DOCNO>
<INTERVIEWDATA> 1926 </INTERVIEWDATA>
<NAME>Chana Lederman</NAME>
<MANUALKEYWORD> Sweden 1945 (January 1 - May 7) | Lund (Sweden) | aid | aid: provision of medical care </MANUALKEYWORD>
<SUMMARY>ZE tells of being taken to a school in Lund, Sweden, where she and her mother were quarantined and given medical care. She recalls receiving gifts from the Swedish population.</SUMMARY>
<ASRTEXT2004A>in and then and uh you can be began they put us on the car why do you recall anything yes they were nurses doctors they checked every day everybody went through eh a checkup they weight loss i was waiting and my mother was waiting forty seven kilos which was like ninety pounds which i did not uh i wait i would more than one i went back to the keep us we stayed there for awhile and they gave us some more calls in the ghetto was and eh this week's came around this was in high school and it was a day why did all of no we'd like your parents it was fenced off they threw wasn't yet standpoint few all my aunt and fruit smoldering all kinds of five she now works to welcome us after awhile in we didn't know nothing we started to gain weight tehran they asked doors here fred we wanna stay which we could war if and we when we go and stopped working in my mind accent i'd rather stopped working their name uhhuh being there is a and this place</ASRTEXT2004A>
<AUTOKEYWORD2004A1> extended family members | occupations, interviewee's | photographs (stills) prewar | education | family businesses | cultural and social activities | fate of loved ones | socioeconomic status | photographs of interviewee (stills) | photographs (stills) 1995 | education in the refugee camps | medical care in the camps | working life in the refugee camps | photographs (stills) 1950s | pregnancies and births | Eisenhower, Dwight D. | Salzkotten (Germany) | Poland 1918 (November 11) - 1939 (August 31) | Lódz (Poland) | Germany 1918 (November 11) - 1933 (January 30) </AUTOKEYWORD2004A1>
</DOC>

Example query

<top>
<num>1159
<title>Child survivors in Sweden
<desc>Describe survival mechanisms of children born in 1930-1933 who spend the war in concentration camps or in hiding and who presently live in Sweden.
<narr>The relevant material should describe the circumstances and inner resources of the surviving children. The relevant material also describes how the wartime experience affected their post-war adult life.
</top>

Weighting schemes for documents and queries (xxx.xxx)

- **Term frequency component**

none (n) : $new\_tf = tf$

max-norm (m) : $new\_tf = \frac{tf}{max\_tf}$

augmented normalized (a): $new\_tf = 0.5 + 0.5 * \frac{tf}{max\_tf}$

log (l): $new\_tf = \ln(tf) + 1.0$

square (s): $new\_tf = tf^2$

- **Merging of collection frequency component**

none (n): $new\_wt = new\_tf$

inverse document frequency weight (t): $new\_wt = new\_tf * \log \frac{N}{df}$

probabilistic (p): $new\_wt = new\_tf * \log \frac{N - df}{df}$

squared (s)

- **Merging of vector normalization**

none (n): $norm\_wt = new\_wt$

sum (s): $norm\_wt = \frac{tf}{\sum_m new\_wt}$

cosine (c): $norm\_wt = \frac{tf}{\sqrt{\sum_m new\_wt^2}}$

Comparison of weighting schemes (Smart, 25 test queries)

	Weighting scheme	TDN	TD	T
		Map	Map	Map
1	lnn.ntn	**0.1366**	**0.1313**	0.1207
2	lnc.ntn	0.1362	0.1214	0.1094
3	mpc.ntn	0.1283	0.1219	0.1107
4	npc.ntn	0.1283	0.1219	0.1107
5	lsn.ntn	0.1195	0.1233	**0.1227**
6	lsn.atn	0.0919	0.1115	**0.1227**
7	nps.ntn	0.0517	0.0416	0.0474
8	mtc.atc	0.1138	0.1151	0.1108

Phonetic transcripts

- The documents and the queries were transcribed in phonetic form and split into 4-grams.
  - NIST's text-to-phone tool http://www.nist.gov/speech/tools/

- Example:
  - child survivors in Sweden
  - ch_ay_l_d s_ax_r_v ax_r_v_ay r_v_ay_v v_ay_v_ax ay_v_ax_r v_ax_r_z ih_n s_w_iy_d w_iy_d_ax iy_d_ax_n

Results on phonetic n-grams, and combination text plus phonetic n-grams (lnn.ntn)

Map	Fields	Description
0.0986	T	Phonetic
0.1019	TD	Phonetic
0.0981	T	Phonetic+Text
0.1066	TD	Phonetic+Text

Results of indexing manual keywords and summaries
(lnn.ntn)

System	Map	Fields
Our system	**0.3256**	TDN
UMaryland	0.3129	TD
Our system	0.2989	TD
Our system	0.2754	T

MAP scores for Terrier and SMART, with and without relevance feedback

	System	63 training queries			42 test queries		
		TDN	TD	T	TDN	TD	T
1	SMART	0.0954	0.0906	0.0873	0.0766	0.0725	0.0759
	SMARTnsp	0.0923	0.0901	0.0870	0.0768	0.0754	0.0769
2	Terrier	0.0913	0.0834	0.0760	0.0651	0.0560	0.0656
	TerrierKL	0.0915	0.0952	0.0906	0.0654	0.0565	0.0685

CLEF 2006 data (63 training queries = 38 +25 queries from CLEF 2005)

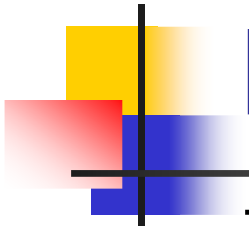
Conclusion

- CLEF 2005: best results out of 7 participating systems,
  CLEF 2006: second place.
- Results improved with:
  - Choice of weighting scheme and fields to index
  - Adding the manual summaries and keywords
- Loss due to Speech Recognition errors

- We can reduce the cWER by up to 50%, by eliminating semantic outliers in:
  - Keyphrases extracted from speech transcripts.
  - Automatic speech transcripts.
- There is a loss of good keywords/content words.

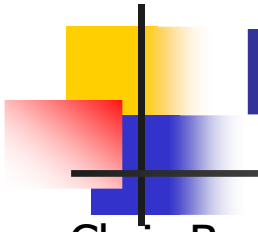
Future work

- Alternative ways to use semantic similarity scores.
- Use clustering for detecting semantic outliers in keywords.
- Use lexical chains for detecting outliers directly in speech transcripts. Treat named entities separately.
- Evaluate on other data sets.
- Evaluate with users the navigation efficiency using filtered keywords and filtered transcripts.
- Future work in IR
  - Filter out potential speech errors – semantic outliers with low PMI score (in a large Web corpus) with neighboring words.
  - Index using speech lattices



References

- Turney, P. D. (2000). Learning algorithms for keyphrase extraction, *Information Retrieval*, 2 (4), 303-336.
- Désilets, A. and de Brujin, B. and Martin. J. (2001). Extracting keyphrases from spoken audio documents. SIGIR Workshop on Information Retrieval Techniques for Speech Applications, 36-50.
- Clarke, C. and Terra, E. (2003). Passage retrieval vs. document retrieval for factoid question answering. ACM SIGIR'03, 327-328.
- Jarmasz, M. and Barrière, C. (2004). Keyphrase extraction: enhancing lists. Proceedings of CLINE'04.
- Diana Inkpen and Alain Desilets, (2004) Extracting semantically-coherent keyphrases from speech, Canadian Acoustics 32(3):130-131, special issue of Acoustics Week in Canada.



References

- Chris Buckley, Gerard Salton, and James Allan. 1993. Automatic retrieval with locality information using SMART. In Proceedings of the First Text REtrieval Conference (TREC-1), p. 59–72.
- Charles L. A. Clarke. 2005. Waterloo Experiments for the CLEF05 SDR Track, in Working Notes for the CLEF 2005 Workshop, Vienna, Austria
- John S. Garofolo, Cedric G.P. Auzanne and Ellen M. Voorhees. 2000. The TREC Spoken Document Retrieval Track: A Success Story. In Proceedings of the RIAO Conference: Content-Based Multimedia Information Access, Paris, France, p. 1-20.
- Douglas W. Oard, Dagobert Soergel, David Doermann, Xiaoli Huang, G. Craig Murray, Jianqiang Wang, Bhuvana Ramabhadran, Martin Franz and Samuel Gustman. 2004. Building an Information Retrieval Test Collection for Spontaneous Conversational Speech, in Proceedings of SIGIR, p. 41-48.
- Gerard Salton and Chris Buckley. 1988. Term-weighting approaches in automatic retrieval. Information Processing and Management, 24(5):513-523.
- Ryen W. White, Douglas W. Oard, Gareth J. F. Jones, Dagobert Soergel and Xiaoli Huang. 2005. Overview of the CLEF-2005 Cross-Language Speech Retrieval Track, in Working Notes for the CLEF 2005 Workshop, Vienna, Austria