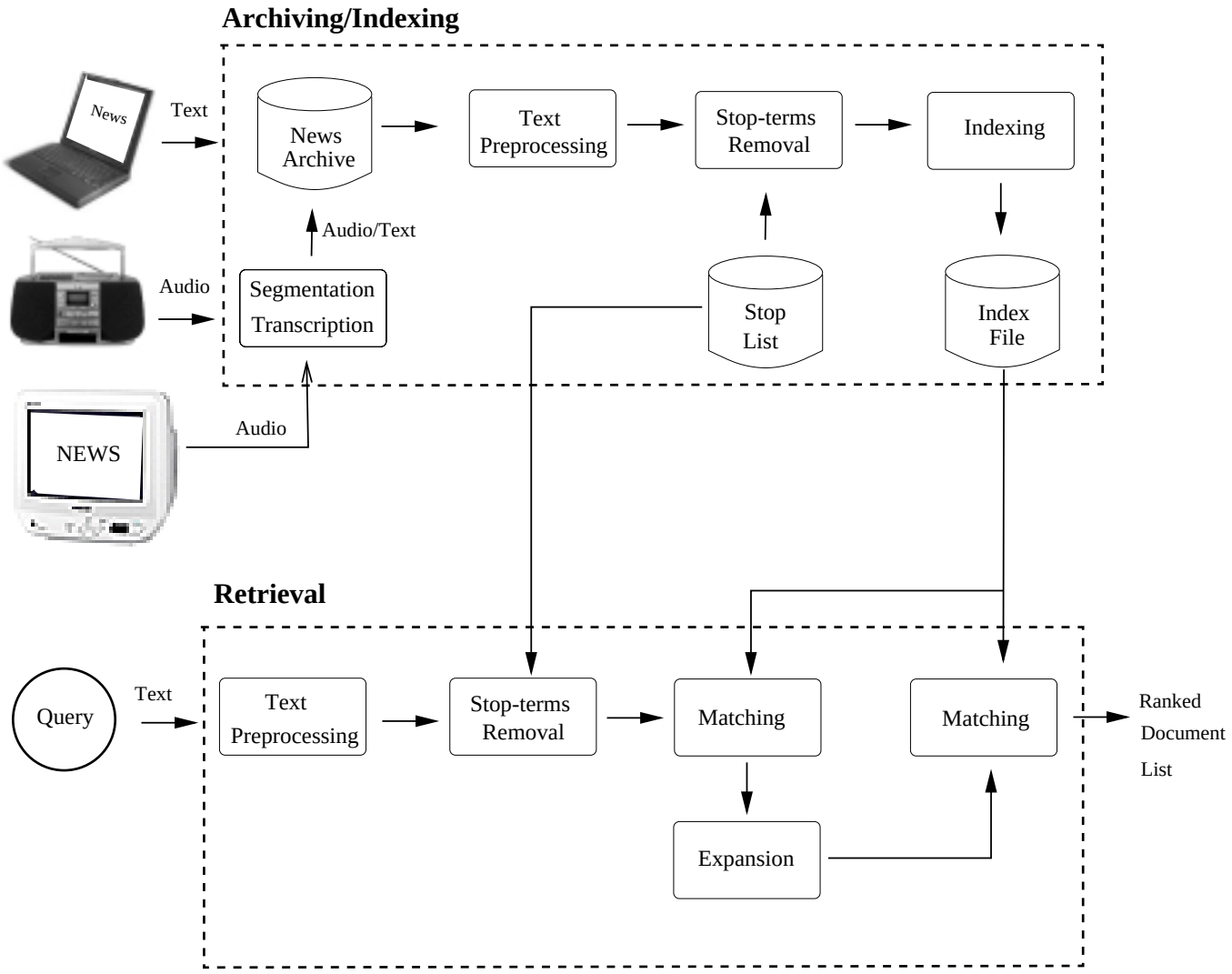
Italian Document Retrieval at ITC-irst

Marcello Federico
Nicola Bertoldi

ITC-irst
Istituto per la Ricerca Scientifica e Tecnologica
I-38050 Povo (Trento), Italy

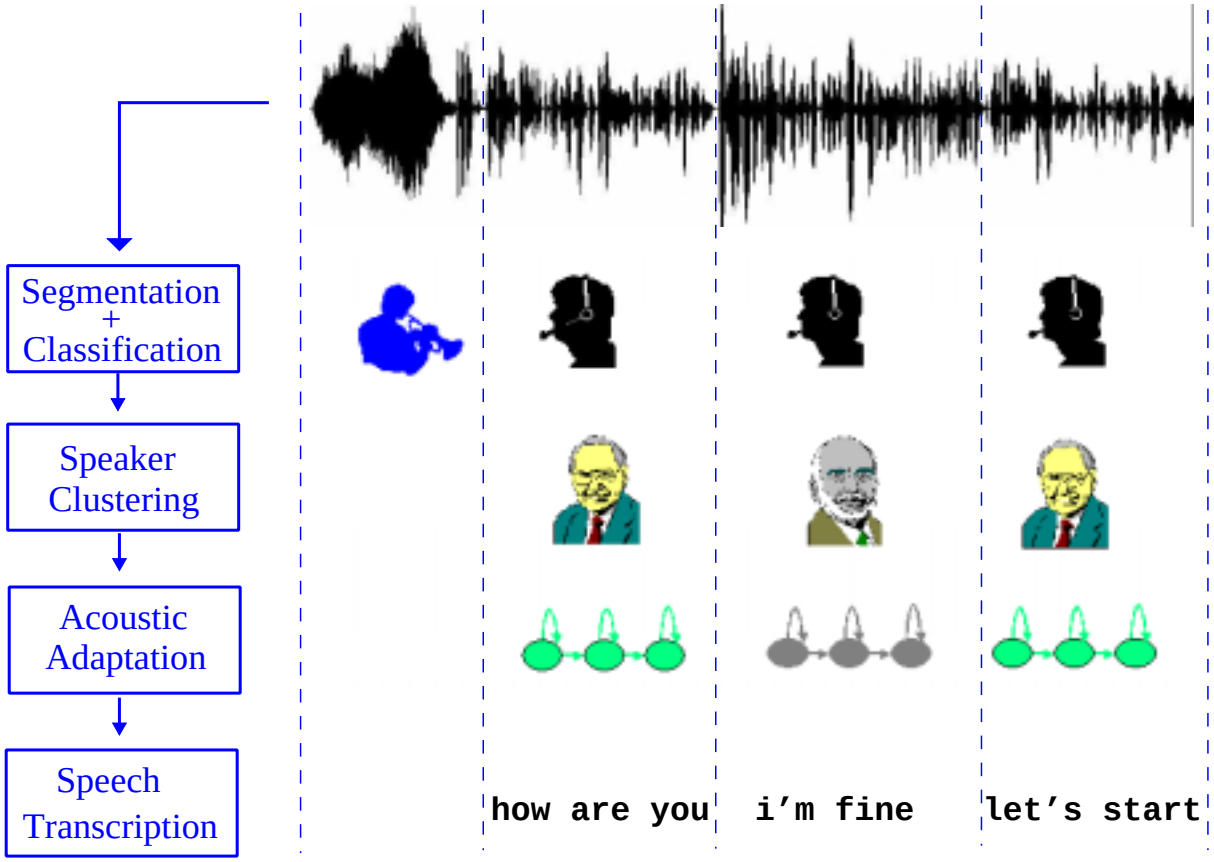
CLEF Workshop - Lisbon, September 22, 2000

Document Retrieval Architecture



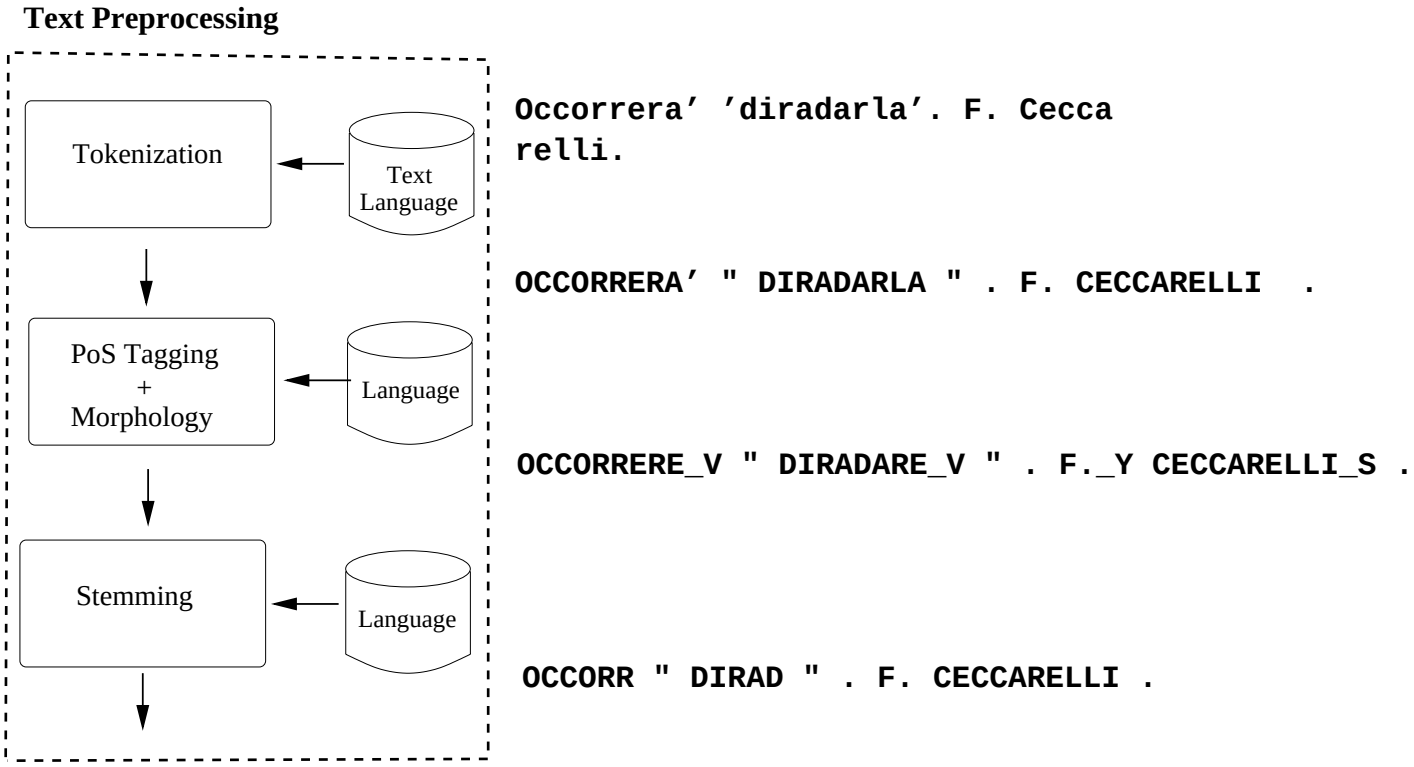
Broadcast News Processing

- CHMMs, CD AMs, 64kw 3gr LM, MLLR adapt
- 20% global WER, 15% on clean speech
- ≈ 20 x real-time (P3 650MHz) - 1Gb RAM



Document Preprocessing

The very task and language-dependent part of the architecture:



Document Matching

Okapi weighting function

$$
\begin{aligned}
Okapi(d,q) &= \sum_{w \in q \cap d} f_q(w)\, c_d(w)\, idf(w) \\
c_d(w) &= \frac{f_d(w)(k_1+1)}{k_1(1-b) + k_1 b \frac{f_d}{l} + f_d(w)} \\
idf(w) &= \log \frac{N - N_w + 0.5}{N_w + 0.5}
\end{aligned}
$$

$f_d(w)$	frequency of word w in document d
$f_q(w)$	frequency of w in query q
f_d	length of document d
$\bar{l}$	mean document length
N	number of documents
N_w	number of documents containing w

Document Matching

Language Model

The query-document matching score is the multinomial log-likelihood:

$$\log P(q \mid d) = \sum_{w \in q} f_q(w) \ \log \ P(w \mid d)$$

The probability of w conditioned to d is estimated through a linear interpolation between the relative frequencies in d and the collection:

$$P(w \mid d) \ = \ \frac{f_d(w)}{f_d + V_d} + \frac{V_d}{f_d + V_d} P(w)$$

$$P(w) \ = \ \frac{f(w)}{f + V} + \frac{V}{f + V} \frac{1}{V}$$

$f_d(w)$	frequency of word w in document d
$f_q(w)$	frequency of w in query q
$f(w)$	frequency of w in the collection
f_d	length of document d
f	length of the collection
V_d	vocabulary size of document d
V	vocabulary size of the collection

CLEF

Development

- Collection: 1999 TREC-8 CLIR track
  (available just few days before the deadline!)
- Tuning: Okapi and Blind Relevance Feedback

Evaluation

- Collection: CLEF 2000 Italian monolingual track
- Official runs (mean average precision):
  Okapi 49% LM 47.5%
- Improvement: Okapi+LM 50.0%

CONCLUSION

Future Work

- Improvement of the statistical approach
- Spoken Document Indexing/Retrieval
- Spoken Document Classification