ITC-irst at CLEF 2002:
Using N-best query translations for CLIR

M. Federico and N. Bertoldi
ITC-irst - Centro per la Ricerca Scientifica e Tecnologica
38050 Povo (Trento) - Italy

Outline

- Introduction
- Statistical CLIR framework
- Statistical models for CLIR
- Experiments
- Conclusions

ITC-irst at CLEF 2002

- Monolingual Italian track (2nd place)
- English-Italian Bilingual track (4th place)
  - original query translation model
  - tools: Italian POS tagger, morphological analyser
  - resources: bilingual dictionary

Statistical CLIR Approach

“Given a query f in a source language (e.g. French), find relevant documents d in the target language (e.g. English) within a collection $\mathcal{D}$”

We express the relevance of d with respect to f with a probability, which has somehow to be modelled.

Statistical document ranking criterion:

$$\textbf{rank}_{d\in\mathcal{D}} Pr(d \mid \mathrm{f}) = \textbf{rank}_{d\in\mathcal{D}} Pr(\mathrm{f}, d) \qquad \textbf{(1)}$$

Statistical CLIR approach

We decompose the basic CLIR probability:

$$\begin{aligned}
\Pr(\mathbf{f}, d) &= \sum_{\mathbf{e} \in \mathcal{T}(\mathbf{f})} \Pr(\mathbf{f}, \mathbf{e}, d) \\
&\approx \sum_{\mathbf{e} \in \mathcal{T}(\mathbf{f})} \Pr(\mathbf{f}, \mathbf{e}) \Pr(d \mid \mathbf{e}) \\
&= \sum_{\mathbf{e} \in \mathcal{T}(\mathbf{f})} \Pr(\mathbf{f}, \mathbf{e}) \frac{\Pr(\mathbf{e}, d)}{\sum_{d'} \Pr(\mathbf{e}, d')}
\end{aligned} \tag{2}$$

- **Assumption:** $\Pr(d \mid \mathbf{f}, \mathbf{e}) = \Pr(d \mid \mathbf{e})$
- **Hidden variable** e **is any translation of** f
- $\mathcal{T}(f)$ **is the set of term-by-term translations of** f

Statistical CLIR approach

$$\Pr(\mathbf{f}, d) \approx \sum_{\mathbf{e} \in \mathcal{T}(\mathbf{f})} \Pr(\mathbf{f}, \mathbf{e}) \frac{\Pr(\mathbf{e}, d)}{\sum_{d'} \Pr(\mathbf{e}, d')} \qquad \textbf{(3)}$$

- $\Pr(\mathbf{f}, \mathrm{e})$ **computed by the query-translation (Q-T) model**
- $\Pr(\mathrm{e}, d)$ **computed by the query-document (Q-D) model**
- **Given that any French term has $\bar{\mathcal{I}} > 1$ translations on average, computation of (3) can be prohibitive:** $O(\bar{\mathcal{I}}^{|\mathbf{f}|})$

Q-D model + Q-T model + approximations ⇒ efficient computation

Query-Document Model

Let $\mathbf{e} = e_1, \ldots, e_n$ and d be a query and a document in English

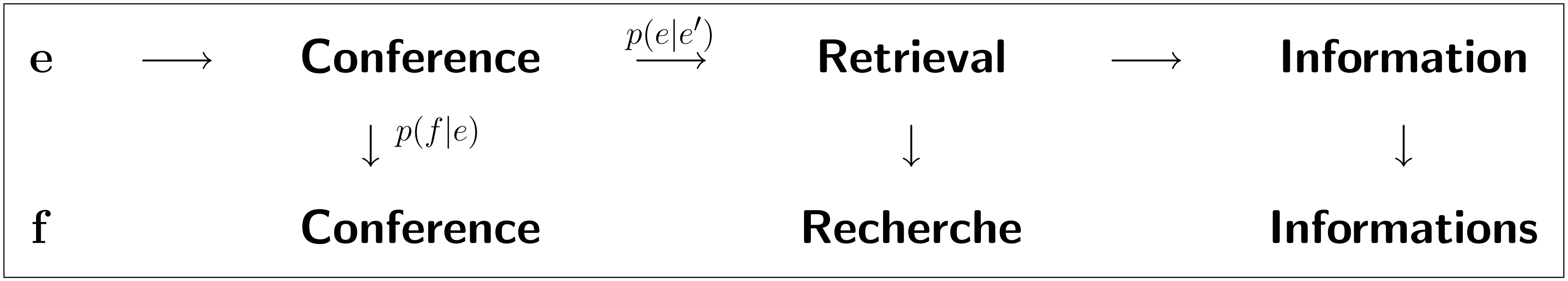
$\Pr(\mathbf{e}, d) = \Pr(\mathbf{e} \mid d)\,\Pr(d)$	**Likelihood x Prior (uniform)**
$\Pr(\mathbf{e} = e_1, \ldots, e_n \mid d) = \prod_{k=1}^{n} p(e_k \mid d)$	**Multinomial model**
$p(e \mid d) = \lambda \frac{N(d,e)}{N(d)} + (1-\lambda)\,p(e)$	**Language Model Smoothing** **(Witten & Bell, 1991)**
$p(e) = \mu \frac{N(e)}{N} + (1-\mu) \frac{1}{\lvert\mathcal{V}\rvert}$	

Query-Translation Model

The translation pair (f,e) is modelled by an Hidden Markov Model:

$$\Pr(\mathbf{f}=f_1,\ldots,f_n,\mathbf{e}=e_1,\ldots,e_n)=p(e_1)\prod_{k=2}^{n}p(e_k\mid e_{k-1})\prod_{k=1}^{n}p(f_k\mid e_k)$$

Dependency graph (or Bayesian network) of a translation:



e $\longrightarrow$ **Conference** $\xrightarrow{p(e|e')}$ **Retrieval** $\longrightarrow$ **Information**

$\downarrow p(f|e)$ $\downarrow$ $\downarrow$

f **Conference** **Recherche** **Informations**

Notice: target LM probs. should cope with word re-orderings!

Estimation of Query-Translation Model

- **Emission/translation probabilities (from bilingual dictionary)**

$$\Pr(f \mid e) = \frac{\delta(f,e)}{\sum_{f'} \delta(f',e)} \qquad \delta(f,e) = 1 \text{ **if** } (f,e) \in \text{ **dict and** } 0 \text{ **otherwise**}$$

- **Transition/target-LM probabilities (from $\mathcal{D}$)**

$$p(e \mid e') = \frac{p(e,e')}{\sum_{e''} p(e'',e')} \qquad p(e,e') = \text{co-occurrence prob in a text window}$$

Smoothing of co-occurrence prob uses non-linear discounting by (Ney et al, 1994), which is suited for symmetric LMs.

Computation with Query-Translation Model

Given a source query f and a Q-T model:

- Viterbi search algorithm permits to efficiently compute the most probable translation:

$$\mathbf{e}^* = \arg \max_{\mathbf{e} \in \mathcal{T}(\mathbf{f})} \Pr(\mathbf{f}, \mathbf{e})$$

- Tree-trellis based search algorithm permits to efficiently compute the N most probable (N-best) translations:

$$\mathcal{T}_N(\mathbf{f}) = \mathbf{e}_1, \mathbf{e}_1, \ldots, \mathbf{e}_N$$

Computation of CLIR Model

As in general few correct translations exist, a reasonable approximation is to marginalize $\Pr(\mathrm{f}, \mathrm{e}, d)$ over the N-best translations:

$$\Pr(\mathbf{f}, d) \approx \sum_{\mathbf{e} \in \mathcal{T}_N(\mathbf{f})} \Pr(\mathbf{f}, \mathbf{e}) \frac{\Pr(\mathbf{e}, d)}{\sum_{d' \in \mathcal{I}(\mathbf{e})} \Pr(\mathbf{e}, d')}$$

- $\mathcal{T}_N(\mathrm{f}) \rightarrow$ **set of N-best translations of** f
- $\mathcal{I}(\mathrm{e}) \rightarrow$ **set of documents containing terms of** e.

CLEF'02 Experimental Evaluation

- **Target collection:**
  - 108,578 Italian articles (La Stampa, SDA - 1994)
- **Test data:**
  - 49 topics (English & Italian)
  - 1072 relevant docs
- **Text Preprocessing:**
  - baseform on Italian and stemming on English
  - Stop-term removal
  - Proper names and number recognition
- **Query Expansion**
  - 15 additional "relevant" search terms are selected from the top 5 docs

CLEF'02 Experimental Evaluation

Official Run	N-best	mAvPr	< mdn	> mdn	wrs	bst
IRSTit1		.4920	11	37	0	7
IRSTen2it1	1	.3444	21	24	3	5
IRSTen2it2	5	.3531	19	26	2	2
IRSTen2it3	10	.3552	16	26	2	6

Other Experiments

Unofficial Run	N-best	mAvPr	Topics
Systran	1	.4037	TD (2nd best score at CLEF 2002)
Systran	1	.4412	TDN
IRST	1	.4285	TDN
IRST	5	.4410	TDN
IRST	10	.4247	TDN