CL-SDR at ITC-irst

Nicola Bertoldi & Marcello Federico

ITC-irst - Centro per la Ricerca Scientifica e Tecnologica

I-38050 Povo (Trento), Italy

{federico,bertoldi}@itc.it

Outline

- Introduction
- Statistical CLIR Model
- Comparative Experiments
- Conclusions

Introduction

"Given a query $\mathbf{f}$ 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 $\mathbf{f}$ with a probability, which has somehow to be modelled.

Statistical document ranking criterion:

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

Statistical CLIR approach

We decompose the basic CLIR probability:

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

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

Statistical CLIR approach

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

- $\Pr(\mathbf{f}, \mathbf{e})$ **computed by the query-translation (Q-T) model**
  - **– defined by an *hidden Markov model* with:**
  - **– emission probs (= lexicon model) estimated from bilingual lexicon**
  - **– transition probs (=target LM) estimated from target collection**
- $\Pr(\mathbf{e}, d)$ **computed by the query-document (Q-D) model**
  - **– defined by a mixture of a SLM and an Okapi model**
- **To speed-up computation of (3) we marginalize over the set of N-best translations of $\mathbf{f}$**
  - **– $\mathcal{T}_N(\mathbf{f})$ is efficiently computed by Viterbi search algorithm + A* algorithm**

(Details in papers at SIGIR 2002 and in special issue of *Information Retrieval*.)

Experiments on CL-SDR

In the following, we present experimental results focusing on:

- **Query translation:**
  - **take 1-best translation from Systran vs.applying the Q-T model**
- **Blind relevance feedback:**
  - **no relevance feedback**
  - **on target collection only**
  - **on parallel collection**
  - **first on parallel collection, then on target collection**
- **Bilingual dictionary:**
  - **freely available vs. commercial (only for Italian)**

Query Translation: Systran vs. Q-T Model

Run	Query	mAvPr
fr-en-1bst-brf-bfr	FR	.2281
fr-en-5bst-brf-bfr	FR	.2314
fr-en-sys-brf-bfr	FR	.3064
de-en-dec-1bst-brf-bfr	DE	.2676
de-en-dec-5bst-brf-bfr	DE	.2660
de-en-sys-brf-bfr	DE	.2880
it-en-1bst-brf-bfr	IT	.2347
it-en-5bst-brf-bfr	IT	.2511
it-en-sys-brf-bfr	IT	.3218
es-en-1bst-brf-bfr	ES	.2746
es-en-5bst-brf-bfr	ES	.2955
es-en-sys-brf-bfr	ES	.3555

Query Expansion

Run	Query	mAvPr
mono	EN	.3176
mono-brf	EN	.3944
mono-brfpar	EN	.3954
mono-brf-brf	EN	.4244
fr-en-5bst	FR	.1555
fr-en-5bst-brf	FR	.2178
fr-en-5bst-brfpar	FR	.2038
fr-en-5bst-brf-bfr	FR	.2314

Bilingual Dictionary

Run	Query	mAvPr
`it-en-5bst-brf-bfr` (free)	IT	.2511
`it-en-col-5bst-brf-bfr` (commercial)	IT	.2648

Conclusions

- Our statistical query-translation model seems less competitive on short queries
  - gap against commercial system slightly reduced by N-best translations
- Query expansion on the contemporary corpus is very helpful
  - maybe other texts can also be helpful
- No difference between free and commercial dictionaries (of comparable size)
  - at least for Italian