TDT4117 - NTNU

# Assignment 2

Aleksander Skraastad
Iver Egge

4. oktober 2013

# 1 - LANGUAGE MODEL

**1a)**
Language models are used in a variety of natural language processing. In speech-recognition-research, probability distributions are buildt that predicts the likelyhood of a given word arising next in a sequence of tokens.
These probability distributions are called language models. In the case of information retrieval, a language model gives the probability that a document would generate the terms of the query, insted of that the query would generate the document.

**1b)**
$q_1$ = *{ NTNU campus }*
$d_1$ = *NTNU is a university in Trondheim*
$d_2$ = *Gløshaugen is a Campus at NTNU, Øya is another campus*

$$p(Q,d) = p(d)\Pi((1-\lambda)p(t) + \lambda p(t|M_d))$$

$$\lambda = \frac{1}{2} = 0.5 = (1-\lambda)$$

For $d_1$

$$p(q_1, d_1) = [(1-\lambda)(\frac{2}{16} + \frac{1}{6})] \times [\lambda(\frac{2}{16} + \frac{0}{6})] = (\frac{7}{192})(1-\lambda)\lambda$$

For $d_2$

$$p(q_1, d_2) = [(1-\lambda)(\frac{2}{16} + \frac{1}{10})] \times [\lambda(\frac{2}{16} + \frac{2}{10})] = (\frac{117}{1600})(1-\lambda)\lambda$$

Therefore

$$d_2 > d_1$$

# 2 - INTERPOLATED PRECISION

**2a)**

Interpolated precision is used when we want to generate a more meaningful graph of recall against precision. Instead of using all the precision values we divide them into precision intervals. We can do this as the user probably wants to retrieve more documents if they are relevant.

Therefore, we maximize each recall-interval, and the graph becomes more ”normalized”.

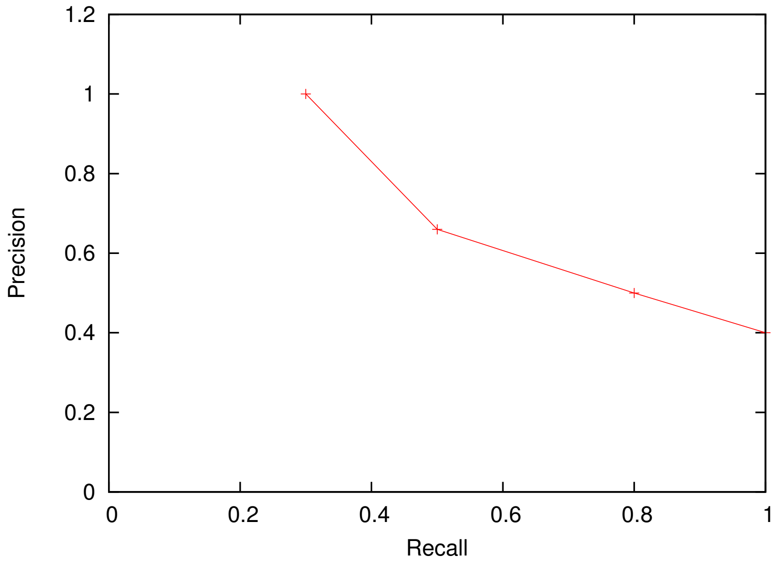
**2b)**

| n | # doc | Relevant | Recall | Precision |
|---|---|---|---|---|
| 1 | 2 | x | $\frac{1}{4}$ | $\frac{1}{1}$ |
| 2 | 64 | | | |
| 3 | 72 | x | $\frac{2}{4}$ | $\frac{2}{3}$ |
| 4 | 10 | | | |
| 5 | 84 | | | |
| 6 | 15 | | | |
| 7 | 103 | x | $\frac{3}{4}$ | $\frac{3}{6}$ |
| 8 | 66 | | | |
| 9 | 37 | | | |
| 10 | 45 | x | $\frac{4}{4}$ | $\frac{4}{10}$ |

Tabell 2.1: Precision and recall

| Standard recall | Precision |
|---|---|
| 0.1 | |
| 0.2 | |
| 0.3 | 1.00 |
| 0.4 | |
| 0.5 | 0.67 |
| 0.6 | |
| 0.7 | |
| 0.8 | 0.50 |
| 0.9 | |
| 1.0 | 0.40 |

Tabell 2.2: Interpolated precision



Figur 2.1: Interpolated Precision

## 3 - Relevance feedback

**3a)**
Relevance feedback is to take the result from an inital query and evaluate if the result is relevant to perform a new query. The evaluation could be done by a user marking the resulting documents as relevant or not.

Query expansion is to process an initial query and improve the formulation of the query or match the query with other, similar queries. This could be to fix spelling errors, matching words in the query with synonyms and/or add to the result other, similar results.

Term reweighting is a mysterious little bastard.

QE and TR were separated at birth.

**3b)**
Automatic global analysis is a technique used to analyse a structured set of texts and look for word relationships (such as word context and phrase structures). Automatic local analysis instead deals with analysis of the documents retrieved by the initial query.

# 4 - Evaluation of IR-systems

**4a)**
Precision and Recall is a measurement of relevance of results. Precision means the fraction of the retrieved result that are relevant, and recall means the relevant fractions of the result. High recall, by this definition, means that most of the relevant information is in the result, while high precision means that there are more relevant than irrelevant information in the result.

Formula for calculating precision

$$precision = \frac{|\{relevantdocuments\} \cap \{retrieveddocuments\}|}{|\{retrieveddocuments\}|}$$

Formula for calculating recall

$$recall = \frac{|\{relevantdocuments\} \cap \{retrieveddocuments\}|}{|\{relevantdocuments\}|}$$

**4b)**
These values are already calculated under "2 - Interpolated precision", task b (ref. table 2.1).