Analysis of Experiments on Hybridization of different approaches in mono and cross-language information retrieval

DAEDALUS – Data, Decisions and Language, S.A.
www.daedalus.es

Universidad Carlos III de Madrid (UC3M)
www.uc3m.es

Universidad Politécnica de Madrid (UPM)
www.upm.es

by José L. Martínez-Fernández

Partially funded by IST-2001-32174 (OmniPaper), CAM 07T/0055/2003 (MIRACLE) and RIMMEL (Multilingual and Multimedia Information Retrieval and its Evaluation) projects

MIRACLE CLEF04 Participation

For CLEF 2004 a total of 62 runs have been submitted for the following tracks:

- **Cross-Language (16)**:

  Experiments described in this presentation.

  - Monolingual **Russian**
  - Monolingual **French**
  - Bilingual **Dutch to French**
  - Bilingual **German to French**

MIRACLE CLEF04 Participation

- **ImageCLEF (45)**:

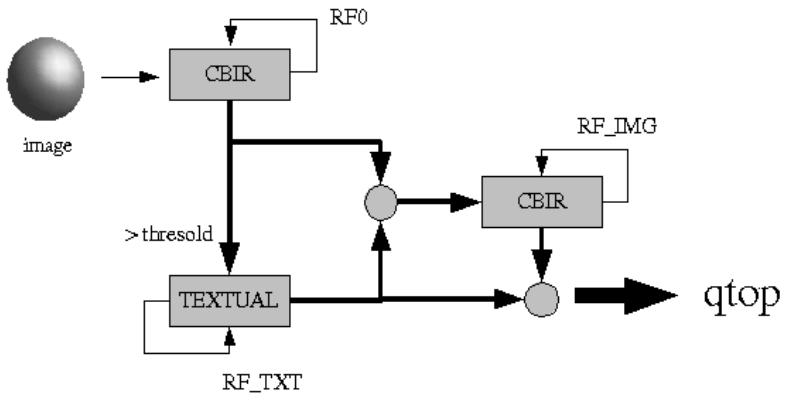
  Text based, image content based and mixing both approaches

  More **linguistic processing** has been applied for English:

  - Query expansion based on **syntactic category**
  - Query translation and expansion using **EuroWordNet**, where possible
  - Influence of **proper noun detection**

  Content based image retrieval

  - Based on GIFT 0.1.9
  - Provided relevance feedback applied

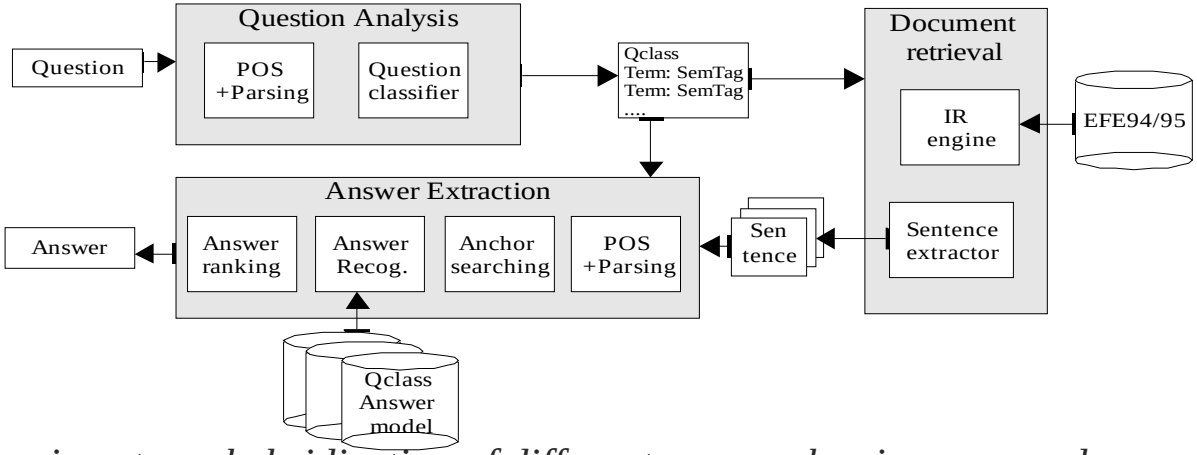


MIRACLE CLEF04 Participation

- **Question Answering for Spanish (1)**

  Our approach has been based on a set of **Markov Models** defined according to predefined question patterns. These Markov Models have been **trained using Google** as a source of data. An **IR system** is previously needed to obtain a small set of documents where the answer can be found.

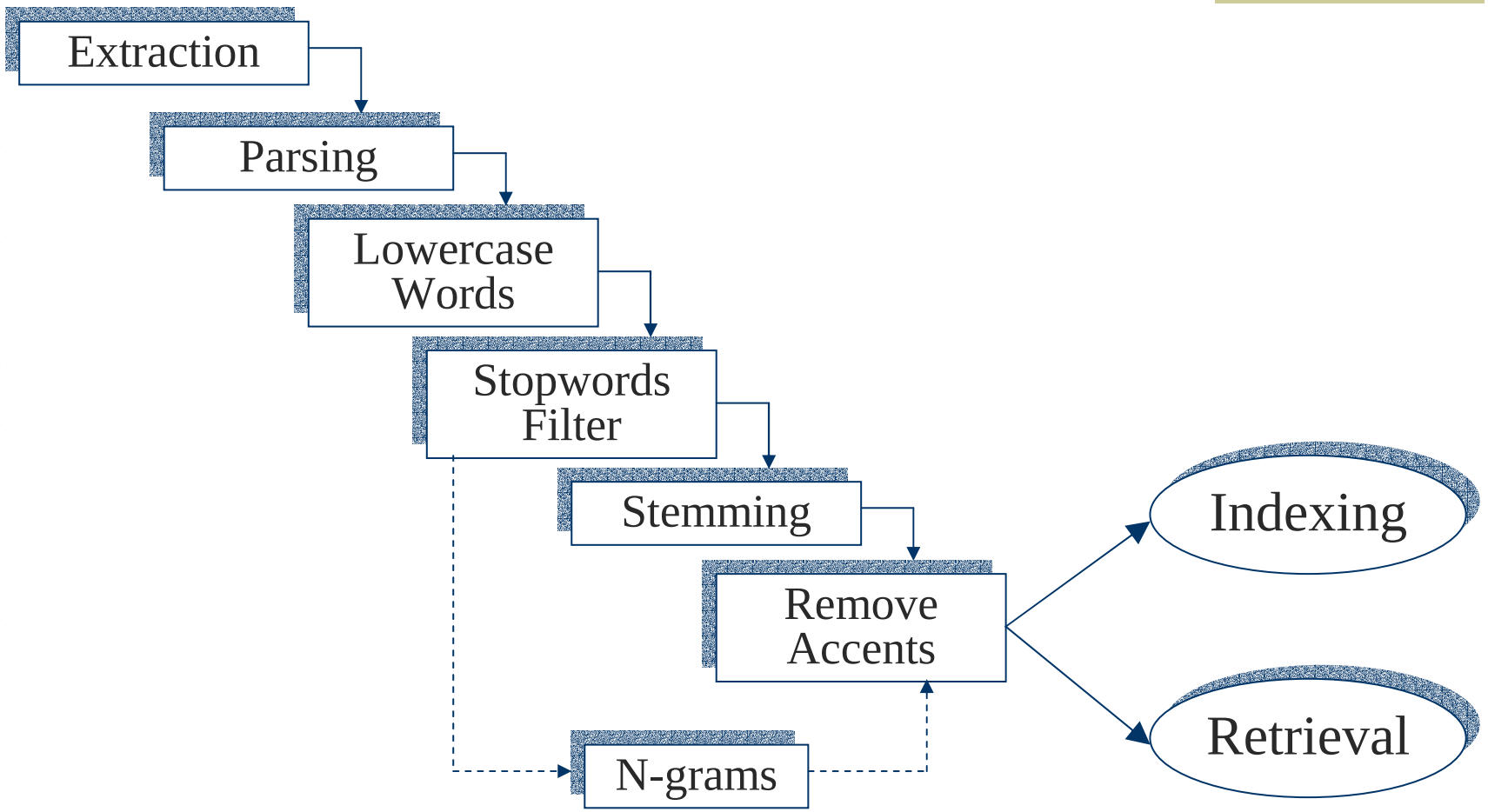
  Only one run, out of contest, could be sent due to some technical problems.



Cross-Language

- Main goal:
  - To **combine** some **basic components** (stemming, transformation, filtering, generation of n-grams, weighting and relevance feedback) in different **structures** and in different **order** of application for document indexing and query processing.
  - A second order combination has been performed, based on **averaging and selective combination of retrieved documents**.

Baseline Approach

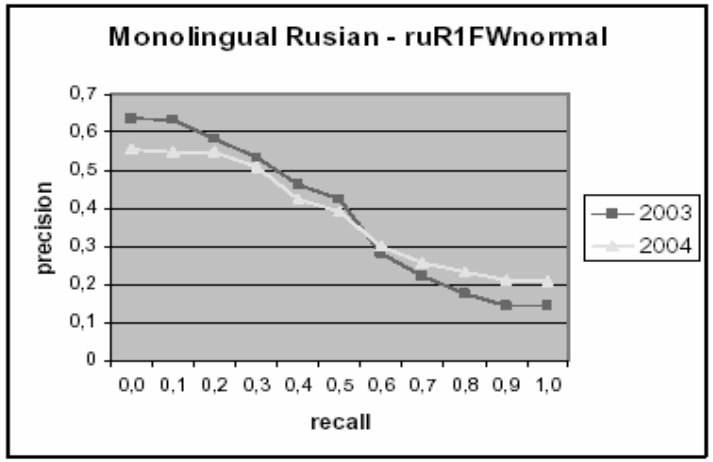


Other techniques applied

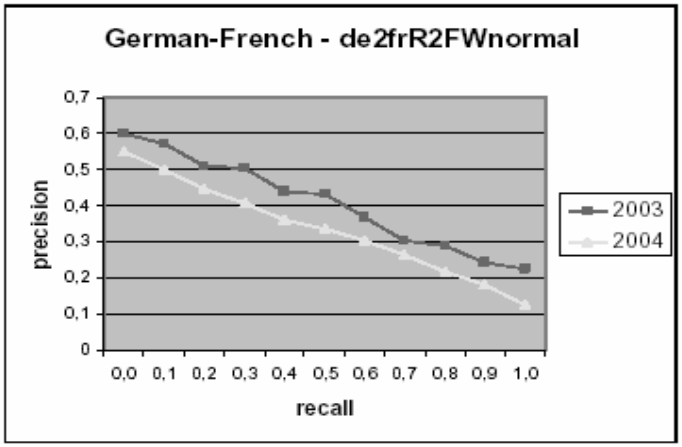
- **Frequent Words**: It consists on filtering out of the queries the 20 most frequent words or stems as well as some typical query terms.
- **Relevance Feedback**: Xapian allows the use of relevance feedback. Several tests made, from 1 to 5 documents used for feedback purposes.
- **Translation tool**: According to results of the tests made in CLEF 2003, the SYSTRAN web translation tool has been used. (For some language pairs not available on-line in SYSTRAN other tools were tested, with very poor results)

Results for Baseline Experiments

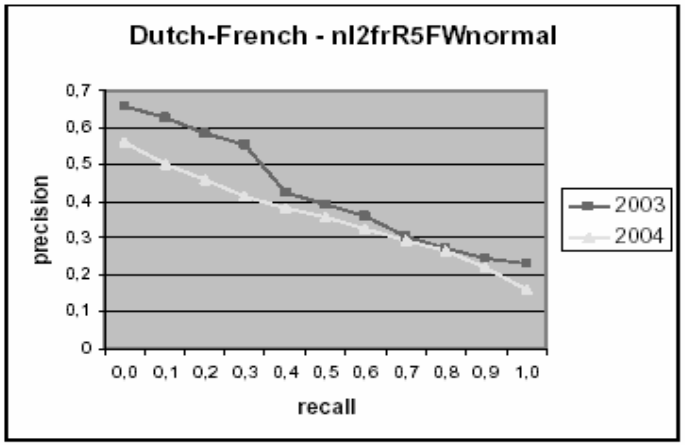
AvgP = 0,3672



AvgP = 0,3201



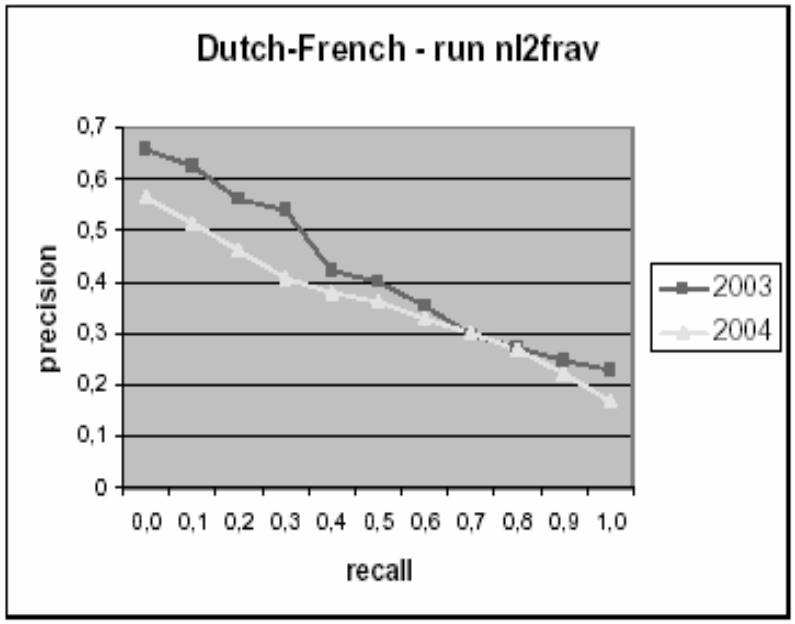
AvgP = 0,3483



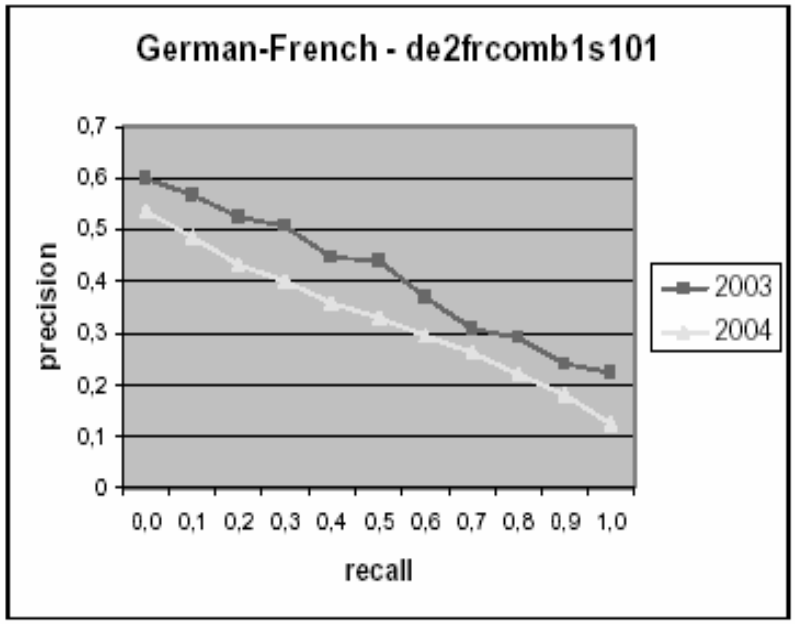
Combined Experiments

- **Combination Methods**
- The idea behind these combinations is that documents having a good score in almost all experiments are more suitable to be relevant than other documents that have good score in one experiment but a bad one in others. Two strategies:
  - **Average:** Relevance results provided by Xapian for a particular document are added. Neither experiment is considered more important.
  - **Asymmetric DWX combination:** The relevant first D documents for each query of the first experiment are preserved for the resulting combined relevance, whereas the relevance for the remaining documents in both experiments are combined using weights W and X. Only “101” and “201” experiments have been ran.

Results for Combined Experiments

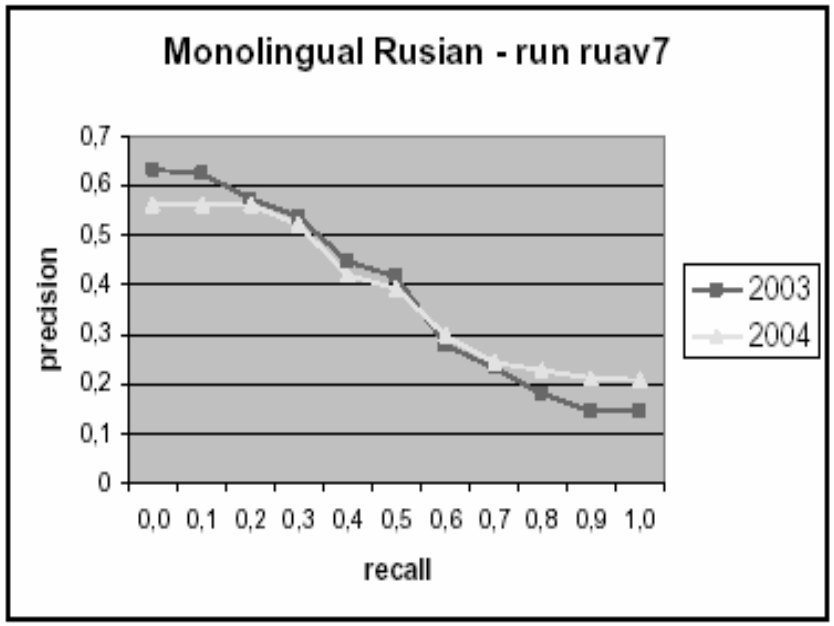


AvgP = 0,3505

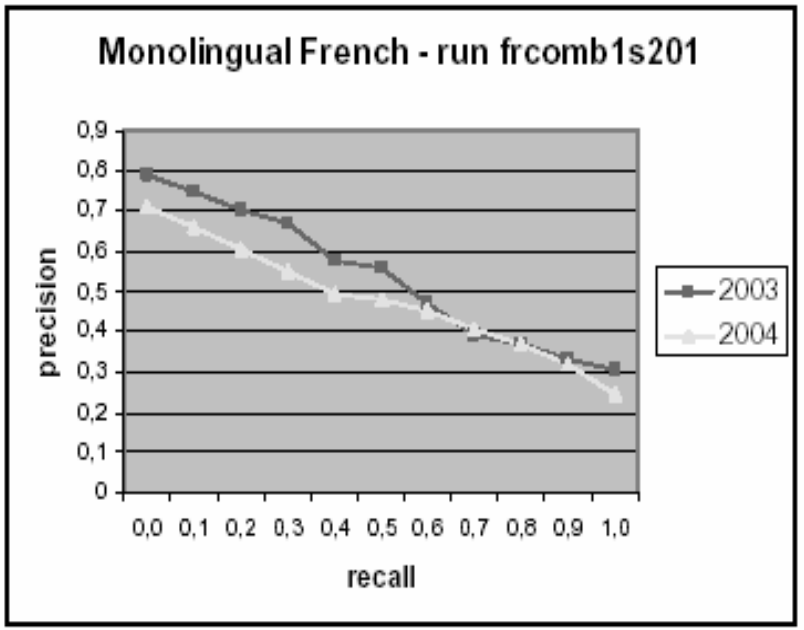


AvgP = 0,3166

Results for Combined Experiments



AvgP = 0,3695



AvgP = 0,4673

Conclusions

- The **combination approach** seems to slightly **improve precision**.
- Differences between 2003 and 2004 data sets show a **dependence on the topic set** supplied each year. In the case of Russian, there is a very low number of relevant documents for the topics set.
- The use of **n-grams** has **not** performed **as expected**. An important decrease in precision is obtained.

Conclusions

- For the basic experiments, conclusions were known in advance, **retrieval performance** can be **improved** by the **use of**:
  - **stemming**
  - **filtering of frequent words**
  - **appropriate weighting**
  - **relevance feedback with a few documents**

Future Work

- Two basic lines:
  - Getting **better performance** in the **indexing** and **retrieval processes**, to be able to make experiments in a more efficient way. This will be done using our own **trie-based libraries** to index and retrieve documents.
  - **Improving** the first **parsing step**, including a good **entity recognition** and **normalization** phase.

The End

THANK YOU FOR YOUR ATTENTION

¿QUESTIONS?