University of Pauda

Department of Information Engineering

DEPARTMENT OF INFORMATION ENGINEERING
UNIVERSITY OF PADOVA

The University of Padua at CLEF 2004: Experiments on Statistical Approaches to Compensate for Limited Linguistic Resources

Giorgio Maria Di Nunzio - Nicola Ferro - Nicola Orio

{dinunzio, nf76, orio}@dei.unipd.it

Workshop of the Cross Language Evaluation Forum 2004 (CLEF 2004)
Bath, UK, 15-17 September 2004

History of the IMS at CLEF

- 2002 Monolingual
  - Language independent stemmer
  - IRON
- 2003 Mono/Bilingual
  - Probabilistic models for automatic stemmer generation
  - Web IRON
- 2004 Mono/Bilingual
  - Limited Language Resources
    - For stemming and query translation
  - IRON enhanced

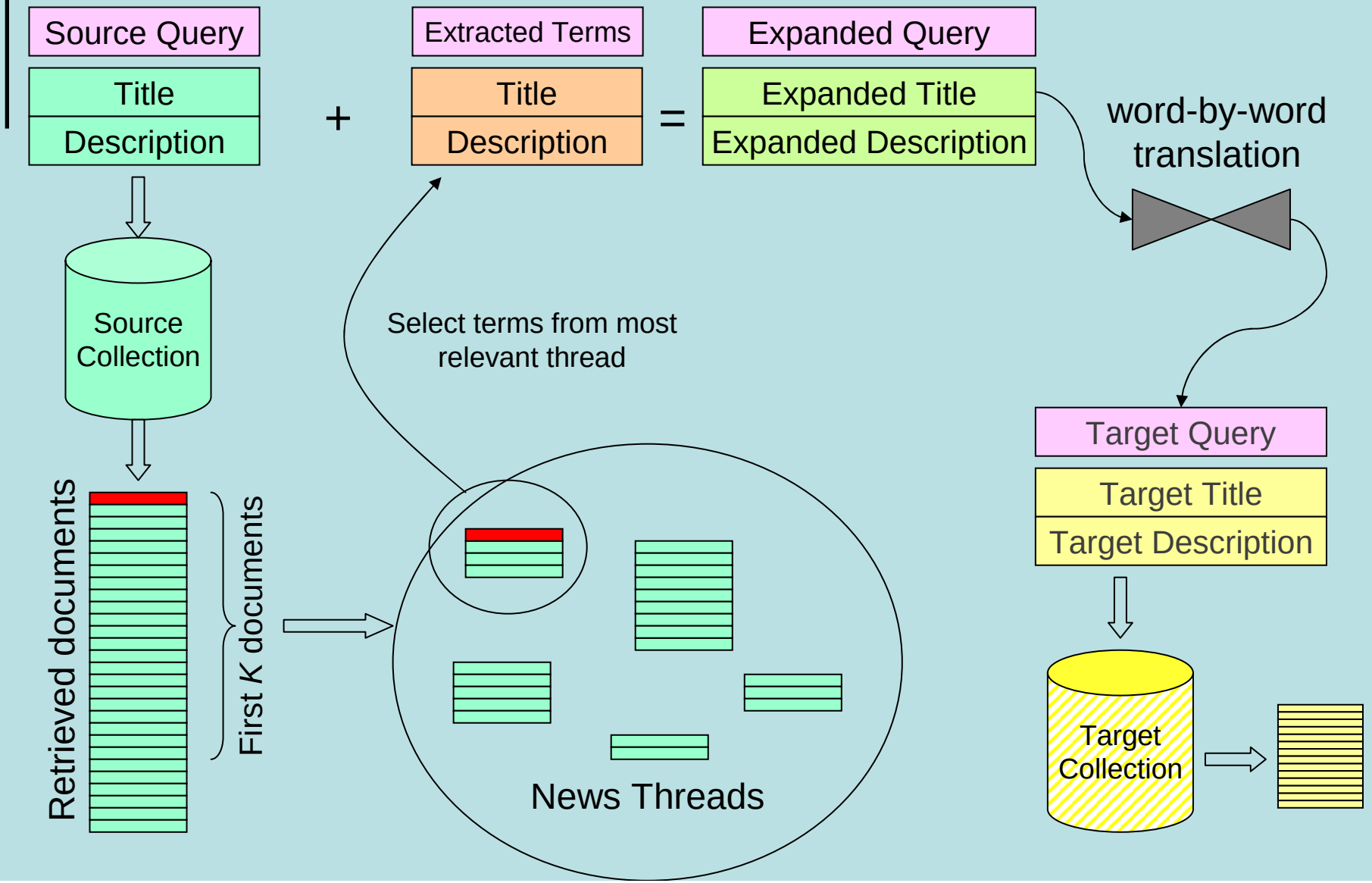
Main Objectives

- Minimize human efforts when applying IR techniques to new languages
- Partially overcome problems of limited language resources
  - Lack of advanced tools for query/document translation
  - Possible lack of knowledge on morphological structure for stemming
- Improve our evaluation prototype system

Bilingual

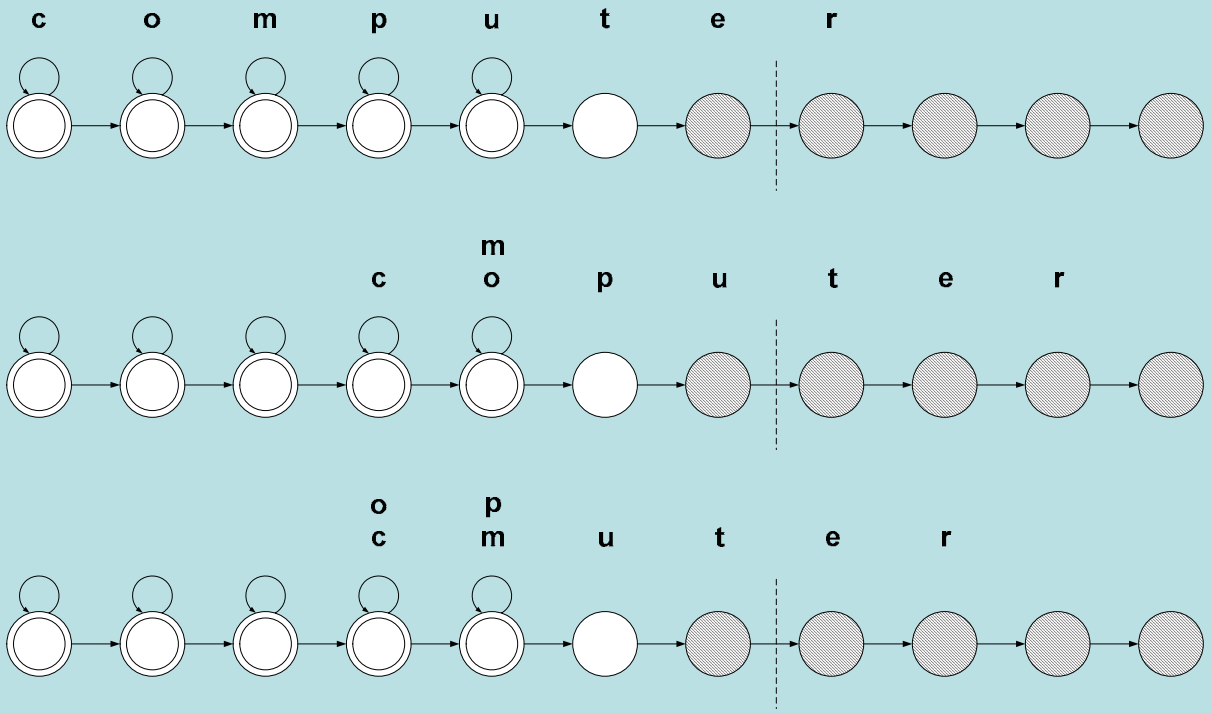
- Almost Comparable Corpora
  - *Automatic news thread identification* by means of hierarchical clustering
- Query expansion
  - Extract significant terms from the most relevant news thread
- Query translation
  - Translate expanded query using on-line word-by-word translation services (Google)
  - No control on the size of the vocabulary
  - No synonyms available

Bilingual

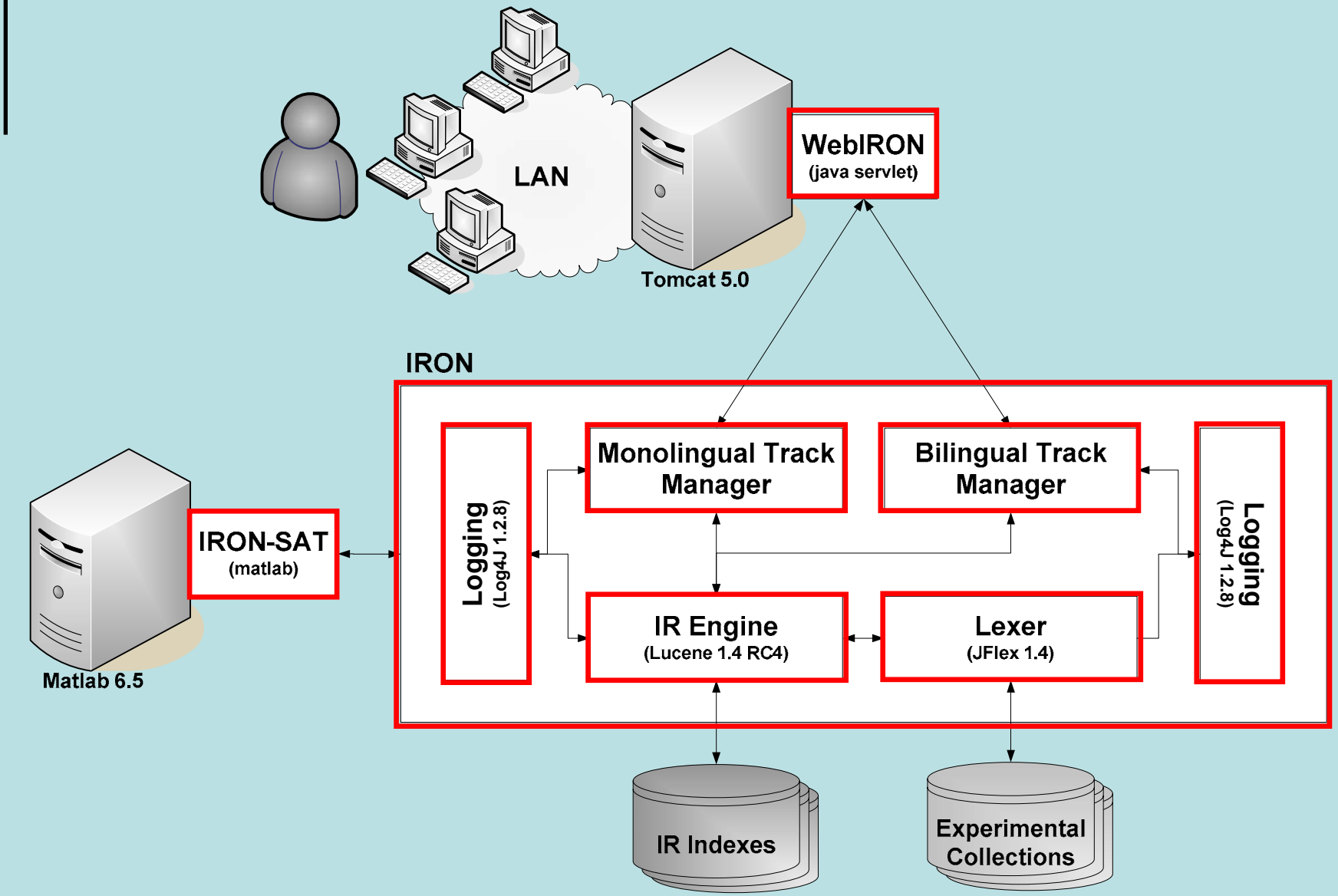

Source Query
Title
Description
+
Extracted Terms
Title
Description
=
Expanded Query
Expanded Title
Expanded Description
word-by-word translation
Source Collection
Select terms from most relevant thread
Target Query
Target Title
Target Description
Retrieved documents
First K documents
News Threads
Target Collection

Monolingual

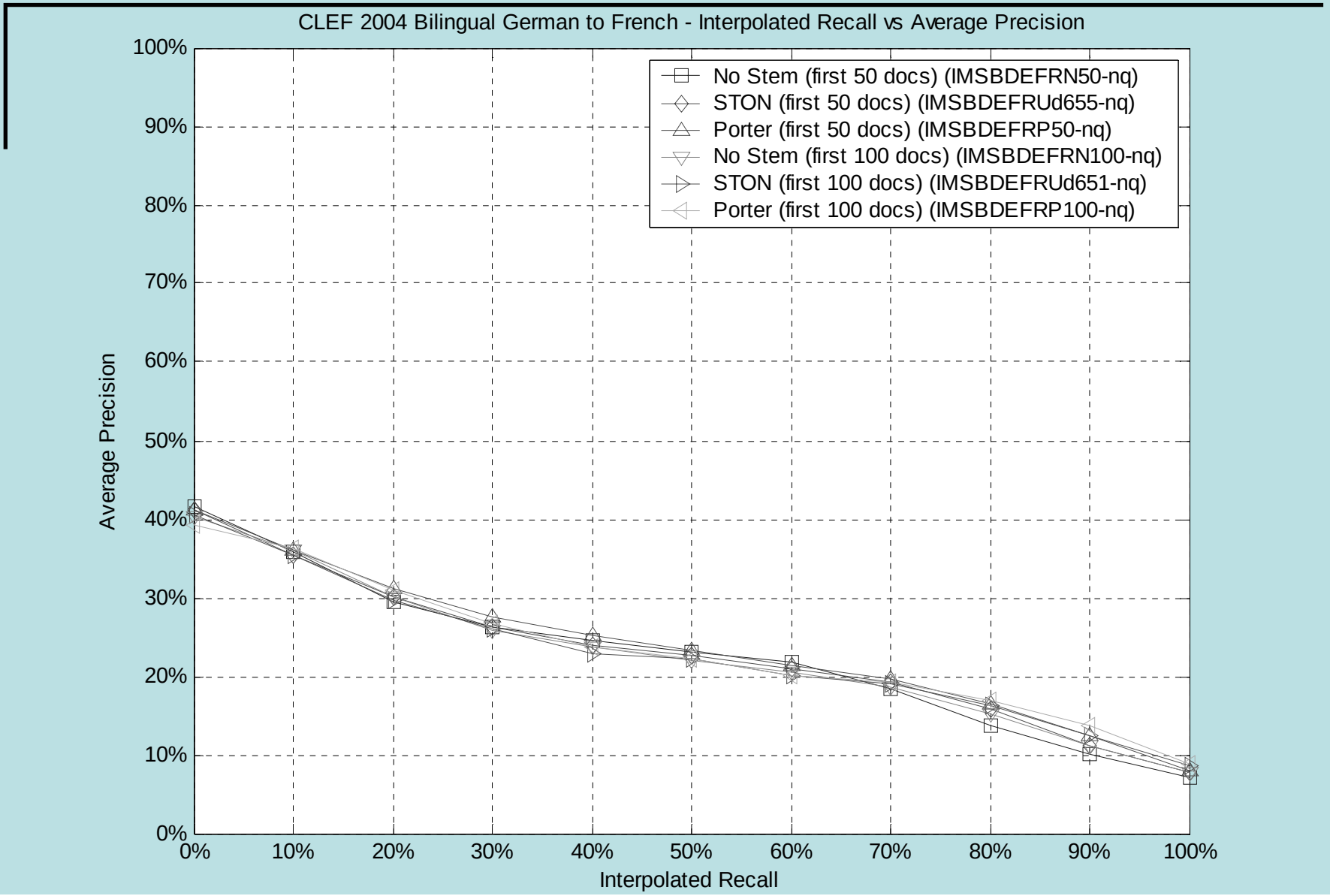
- STON : Hidden Markov Models stemmer
  - The sequence of letters of a word can be considered as a sequence of symbols emitted by a HMM
  - Most probable path for the observed word
  - Transition from stem-set to suffix-set (split-point)
  - Only needs a set of words of the language to be trained off-line



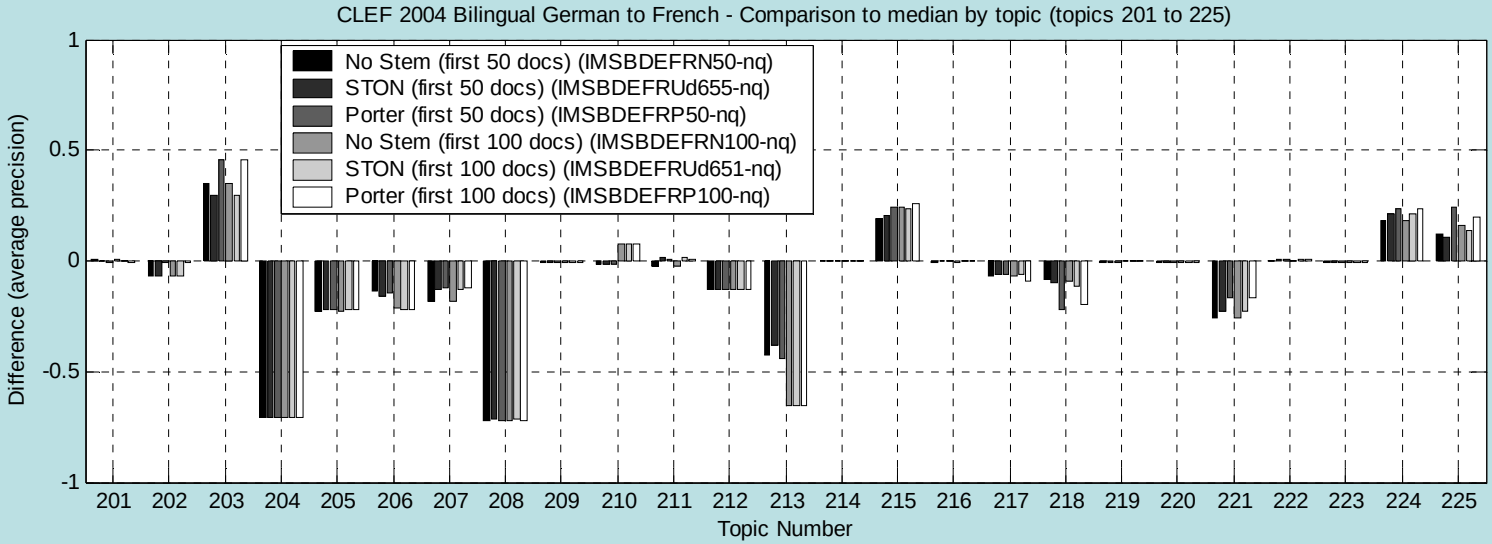
IRON System

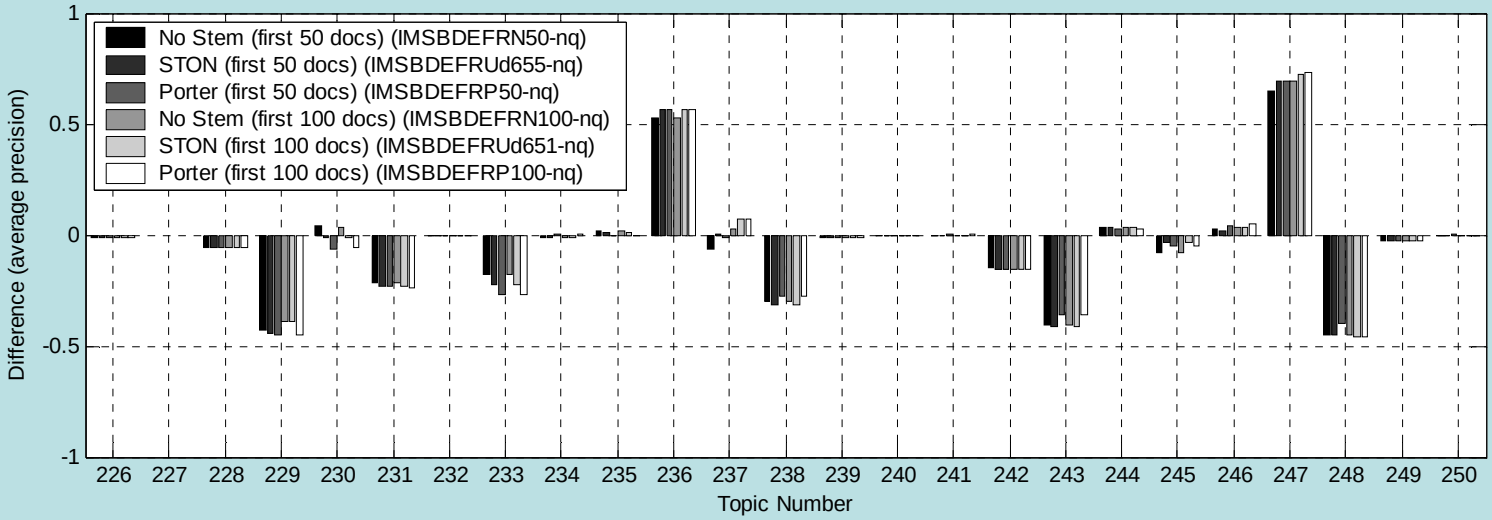

LAN
WebIRON
(java servlet)
Tomcat 5.0
IRON
Logging
(Log4J 1.2.8)
Monolingual Track Manager
Bilingual Track Manager
Logging
(Log4J 1.2.8)
IRON-SAT
(matlab)
Matlab 6.5
IR Engine
(Lucene 1.4 RC4)
Lexer
(JFlex 1.4)
IR Indexes
Experimental Collections

Bilingual Experiments

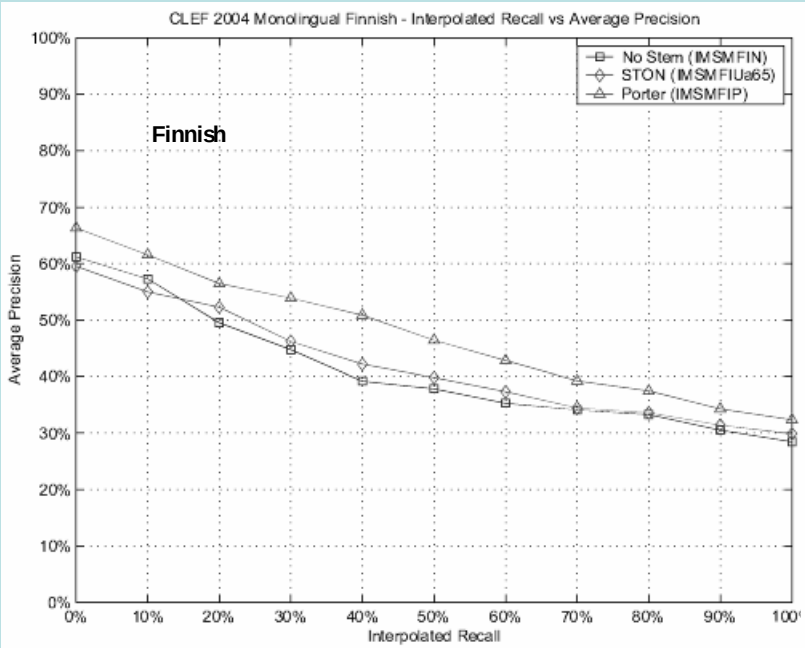


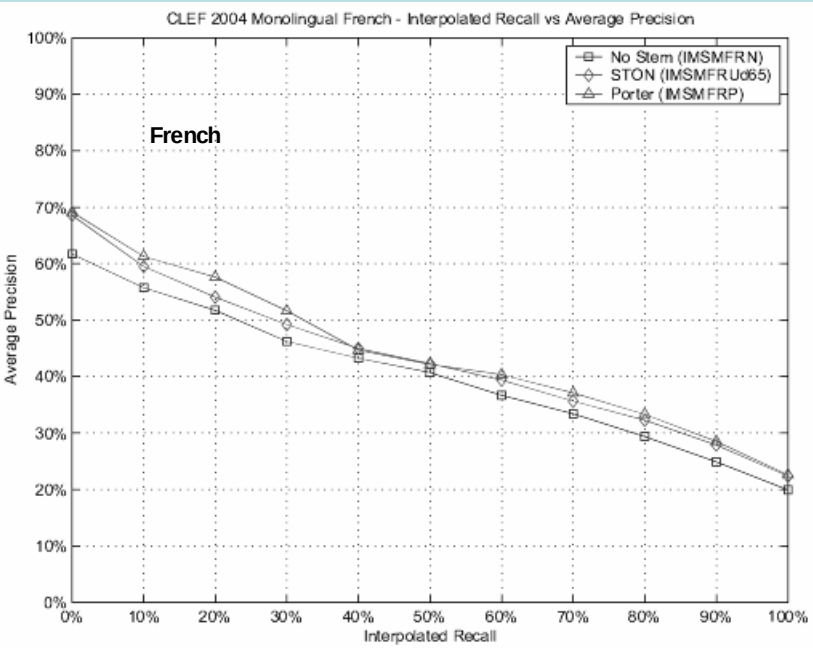
Bilingual Experiments

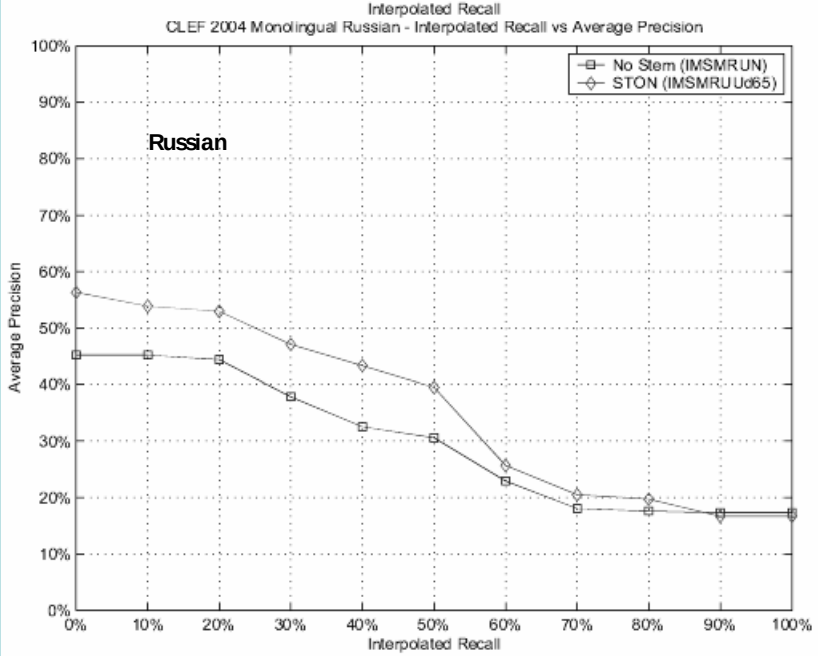




Monolingual Experiments







Comparison STON — No Stem (p-value %)			
	Finnish	**French**	**Russian**
Rel. Retr.	1.93 %	0.54 %	0.39 %
Avg. Prec.	65.84 %	29.52 %	6.79 %
Exact R-Prec.	56.29 %	20.85 %	57.81 %

Conclusions and Future Work

- Minimizing human labor and language resources
  - Automatic stemmer generation
  - Automatic query expansion by means of hierarchical clustering
  - Free on-line word-to-word translation
- Statistical analysis of results

- Thread identification in both source and target collection.
  - Coupling between threads to refine results