N-grams for Translation and Retrieval in CL-SDR

Paul McNamee and James Mayfield

Johns Hopkins University Applied Physics Laboratory
11100 Johns Hopkins Road
Laurel MD 20723-6099 USA
{mcnamee,mayfield}@jhuapl.edu

RTDC
RESEARCH & TECHNOLOGY DEVELOPMENT CENTER

Tokenization in Monolingual SDR

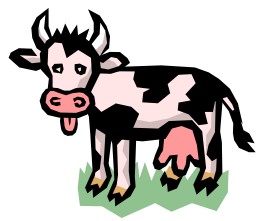
- Use of words / stems is routine
- 'Experiments in Spoken Document Retrieval Using Phoneme N-grams', C. Ng, R. Wilkinson, and J. Zobel, 1998.
  - Generally not as effective as words
  - Combinations of n-grams (e.g., n=3,4,5), beneficial
- 'Subword-based Approaches for Spoken Document Retrieval', K. Ng., PhD Thesis, MIT, 2000.
  - N-grams, n=3 worked as well as text, only if phonemes were accurately recorded
  - Didn't compare against n-grams of transcribed words, but noted it as an interesting area to pursue

N-gram Translations

- Just as words can be statistically translated using an aligned bitext, so can n-grams
  - Difficult to quantify accuracy of mappings
- May mitigate problems in dictionary-based CLIR
  - word lemmatization
  - multiword expressions
  - out of vocabulary words, particularly names
- Hypothesis: N-grams will be robust against ASR errors

	German	Italian
word	milch	latte
stem	milch	latt
4-grams	milc ilch	latt latt
5-grams	_milc milch ilch_	_latt _latt latte

	French	Dutch
word	lait	melk
stem	lait	melk
4-grams	lait	melk
5-grams	_lait lait_	_melk melk_



Translations of Proper Names, Phrases

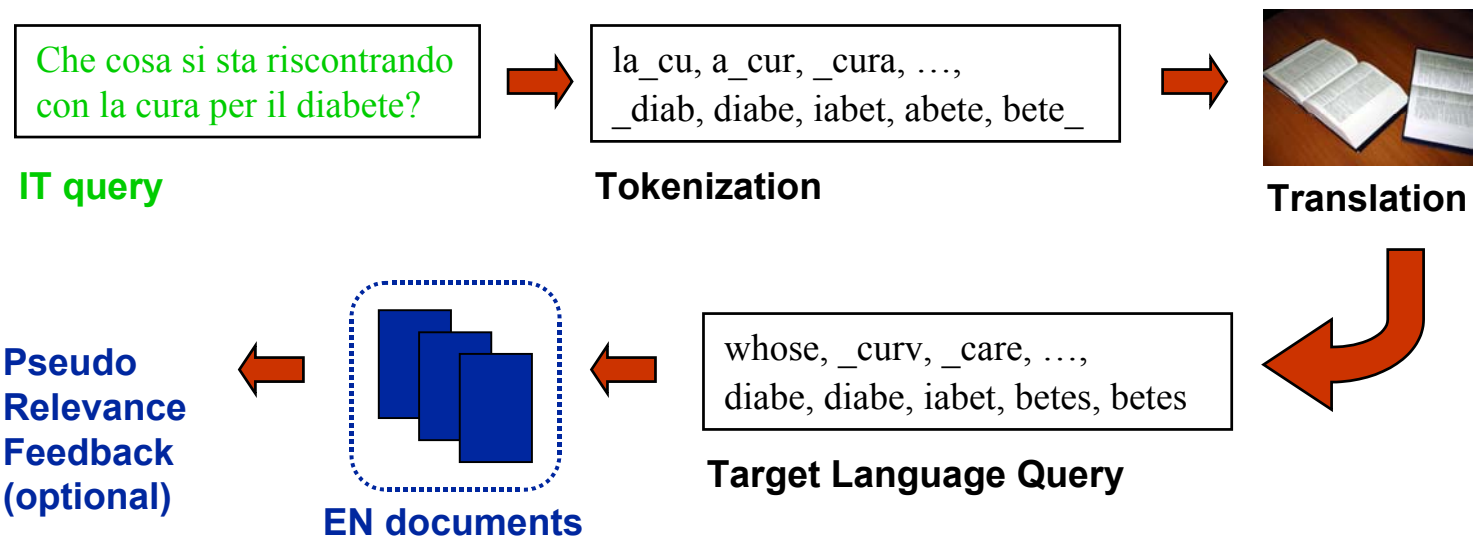
- **French to English:**
  - **Mexique to Mexico (topic 85)**
- **German to English**
  - **Tabakindustrie to 'tobacco industry' (topic 89)**

	French	English
5-grams	_mexi mexiq exiqu xique ique_	_mexi mexic mexic oxic_ ical_

	German	English
5-grams	_taba tabak abaki bakin akind kindu indus ndust dustr ustri strie trie_	bacco bacco bacco cco_i cco_l ustry indus indus indus indus indus indus

N-gram CL-SDR

1. **Indexed Transcribed Target Language Text**
2. **Tokenize Query in Same Fashion as Documents**
3. **One-best Translation of Each Token From Source Language to Target Language**
4. **Relevance Feedback Performed**



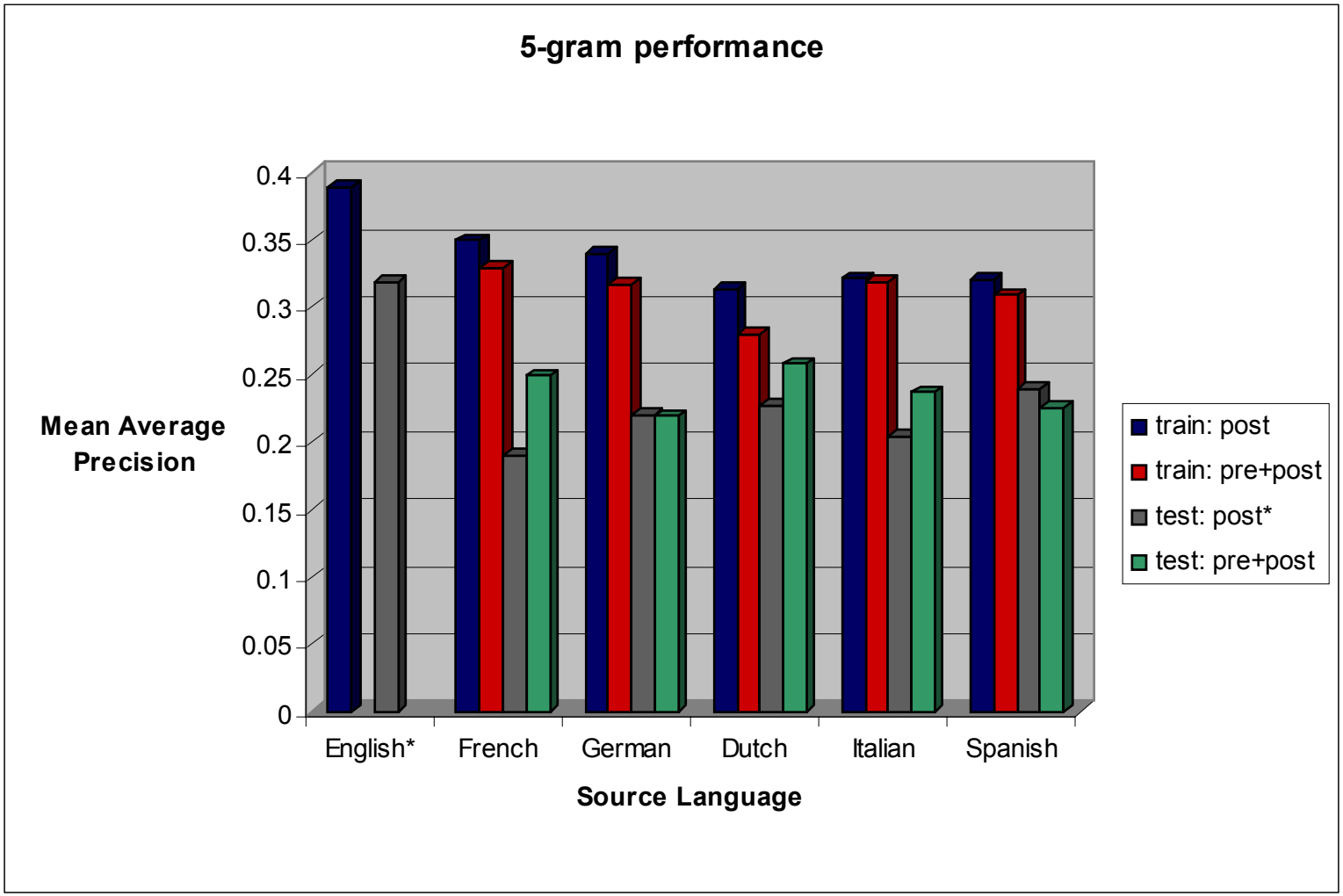
Approach

- **Used the HAIRCUT system**
  - **Java based system described in CLEF 2001 report**
  - **Typically use overlapping character n-grams**
- **Statistical Language Model**
  - **Like Ponte/Croft, Hiemstra/de Vries**
  - **Differs in method for probability estimatation**
- **Relevance Feedback (when used)**
- **Query translation**
- **Post hoc use of CLEF source language collections for pre-translation query expansion**
  - **Translating ~60 terms vs. original query seems to be highly effective**

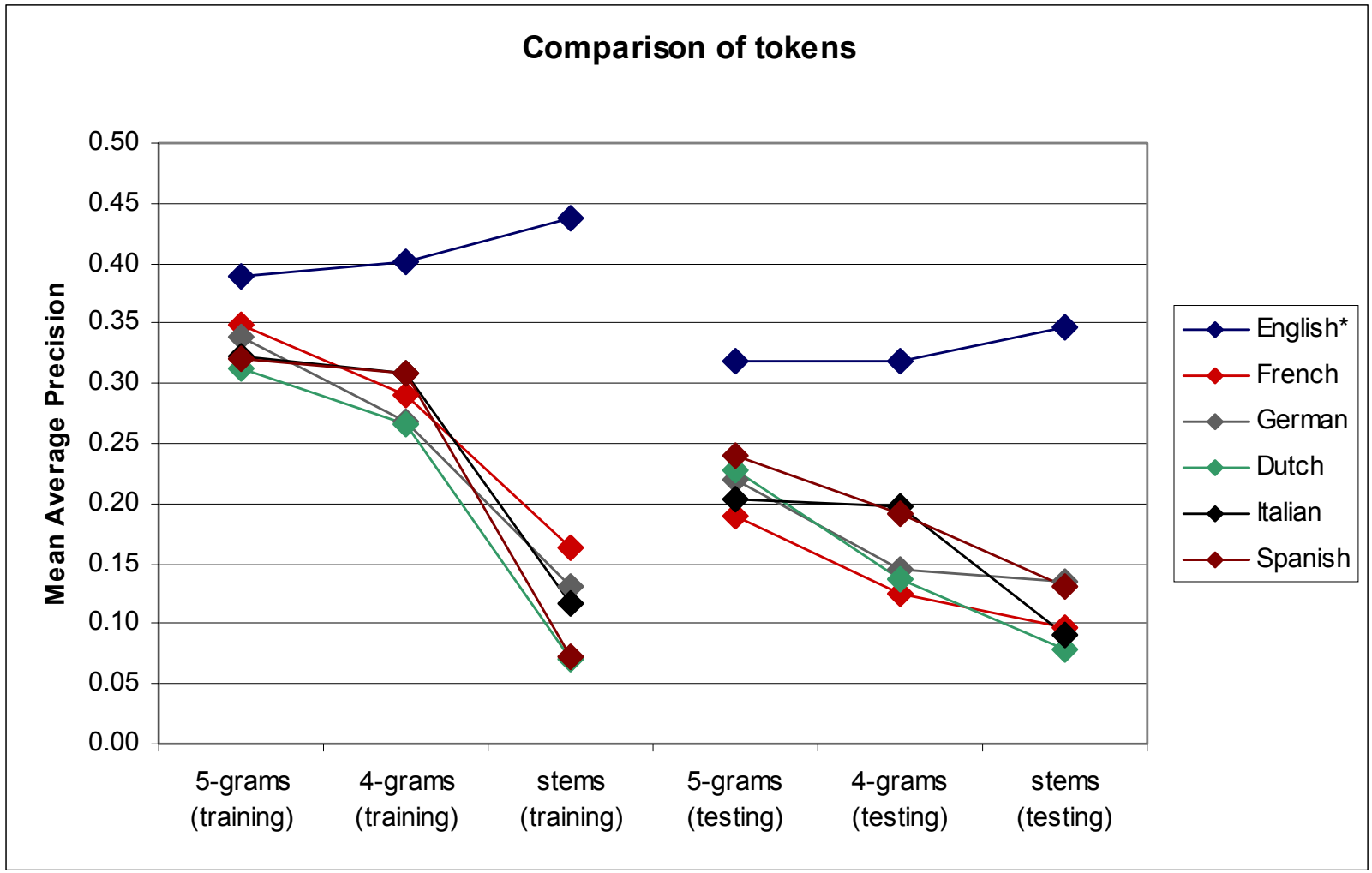
Task

- **Collection**
  - 21738 documents (TREC-8 and TREC-9 SDR Track)
- **HAIRCUT Indexes**
  - Stems (23174 terms; 8 minutes to build)
  - 4-grams (44571 terms; 22 minutes to build)
  - 5-grams (179876 terms; 33 minutes to build)
- **Topics**
  - 50 for training (TREC-8), 50 for test (TREC-9)
- **Training**
  - We didn't really do any
  - We built several indexes and decided on 5-grams with blind relevance feedback

Results


5-gram performance
Mean Average Precision
0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
English*
French
German
Dutch
Italian
Spanish
Source Language
train: post
train: pre+post
test: post*
test: pre+post

Tokenization and CL-SDR



Examination of English Topics

- **Atypical representation of abbreviations was observed post facto**
  - **non standard interior spacing (U. S. vs. U.S.)**
  - **use of periods (P. G. A. vs PGA; H. I. V. vs HIV)**
  - **3/50 training, but 14/50 test topics – an issue for CL-SDR**

- **Proper names in topic sets**
  - **People's names**
    - **Training: 2 topics ('Saddam Hussein' & 'Seinfeld')**
    - **Test data: 7 topics**
  - **Places (below national level, e.g., Boston, Arkansas)**
    - **not including "US", "Europe", "world"**
    - **Training: 16**
    - **Testing: 5**

Summary

- **Encouraged by our first efforts at CL-SDR**
  - **5-gram tokenization and direct 5-gram translation appears to be effective**
- **Puzzling discrepency between effectiveness of tokenization methods in monolingual and CL-SDR**
- **Future issues to explore**
  - **manually normalizing abbreviations**
  - **use of contemporaneous expansion / translation corpora**
  - **use of unstemmed words**