JHU/APL Experiments in Tokenization and Non-Word Translation

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

21 August 2003

RTDC
RESEARCH & TECHNOLOGY DEVELOPMENT CENTER

Previous APL Research at CLEF

- **CLEF 2000**
  - Initial exploration of MT & parallel texts for translation
  - Comparing n-grams (*n*=*6*) and words for retrieval
- **CLEF 2001**
  - Comparing translation resources
  - Score normalization for multilingual merging
  - Examining pre-translation query expansion
- **CLEF 2002**
  - Exploration of no-translation retrieval (n-gram cognates)
  - Translation of document representations (vs. query translation)

CLEF 2003 Motivation

- **Many questions about tokenization remain un/under-addressed**
  - Importance of diacritical marks
  - Variability in performance due to n-gram length
  - Variations across languages
  - Relative efficacy of n-grams and stemmed words
  - Performance implications of n-grams
  - Hybrid methods

- **Tokenization affects Translation**
  - We examined a new method for query translation

General Approach

- Used the HAIRCUT system
  - Java based system described in CLEF 2001 report
- Statistical Language Model
  - Requires one smoothing parameter
  - Differs in method for probability estimation
- Blind Relevance Feedback (optionally)
- Query translation (for bilingual runs)
  - Used CLEF source language collections for pre-translation query expansion to 60 terms
  - Translating a set of ~60 terms vs. original query seems to be highly effective
- Uniform processing for each language

Statistical Language Models

- **HAIRCUT uses a linguistically-motivated probabilistic model to estimate the probability that a document is relevant given a query**
  - **Ponte and Croft, *(SIGIR-98)***
  - **Miller, Leek, and Schwartz, *(SIGIR-99)***
  - **Hiemstra and de Vries, (*CTIT Tech. Report*, May 2000)**

$Q =$ query

$q =$ word in query

$D =$ document

$R =$ set of relevant documents

$\lambda =$ a random Boolean variable

$$P(D \in R \mid Q) = \frac{P(Q \mid D \in R) P(D \in R)}{P(Q)} \quad \text{Bayes law}$$

$$\propto P(Q \mid D \in R) \quad \text{assume constant priors}$$

$$= \prod_{q \in Q} P(q \mid D \in R) \quad \text{Naïve Bayes assumption}$$

$$= \prod_{q \in Q} \left[P(q \mid D \in R, \lambda) P(\lambda) + P(q \mid D \in R, \bar{\lambda}) P(\bar{\lambda}) \right] \quad \text{introduce } \lambda$$

$$= \prod_{q \in Q} \left[\alpha P(q \mid D \in R, \lambda) + (1 - \alpha) P(q \mid D \in R, \bar{\lambda}) \right] \quad \text{define } \alpha = P(\lambda)$$

$$= \prod_{q \in Q} \left[\alpha P(q \mid D \in R, \lambda) + (1 - \alpha) P(q \mid \bar{\lambda}) \right] \quad \text{if } q \text{ ind. of } D \text{ given } \lambda$$

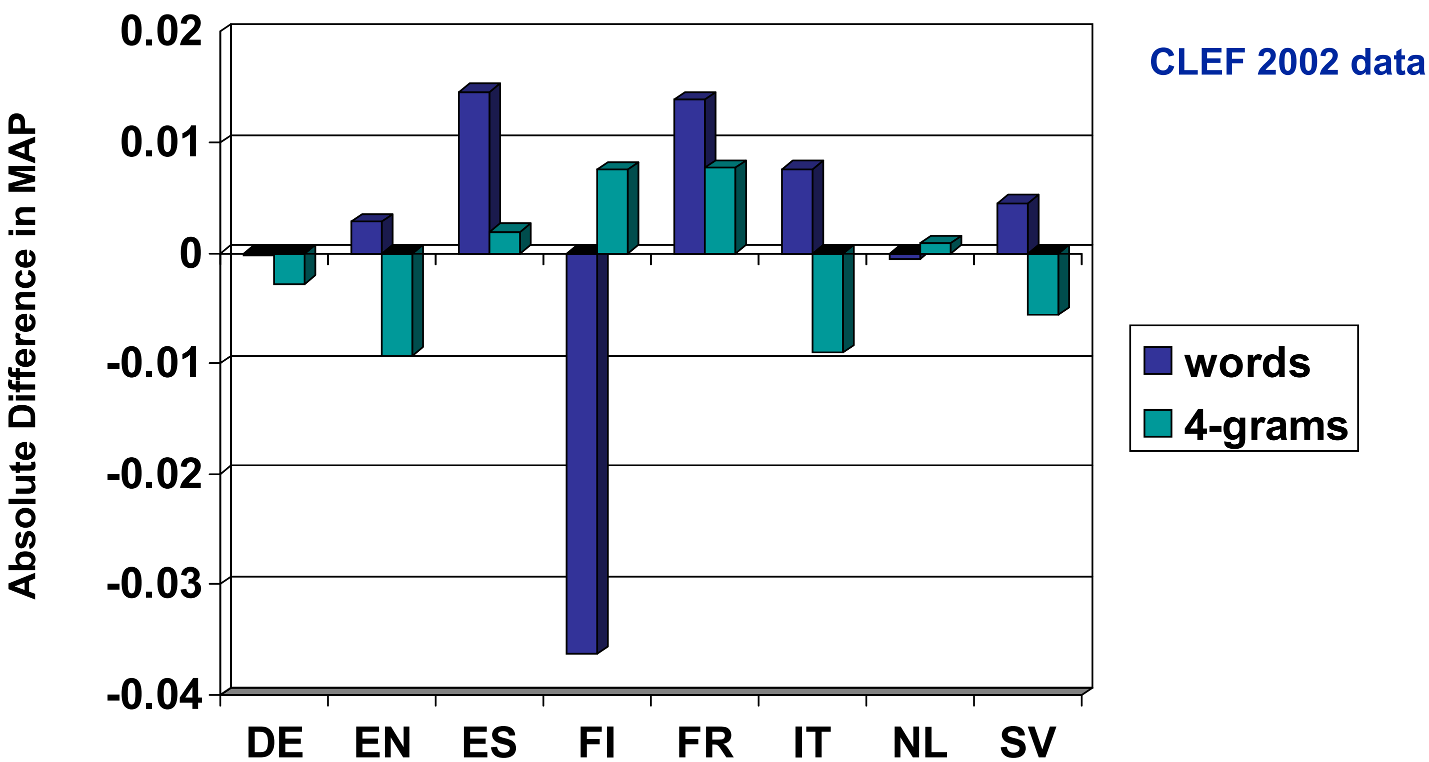
$$= \prod_{q \in Q} \left[\alpha P(q \mid D \in R) + (1 - \alpha) P(q) \right] \quad \text{because lambdas are ugly}$$

relative document term frequency

mean relative document term frequency

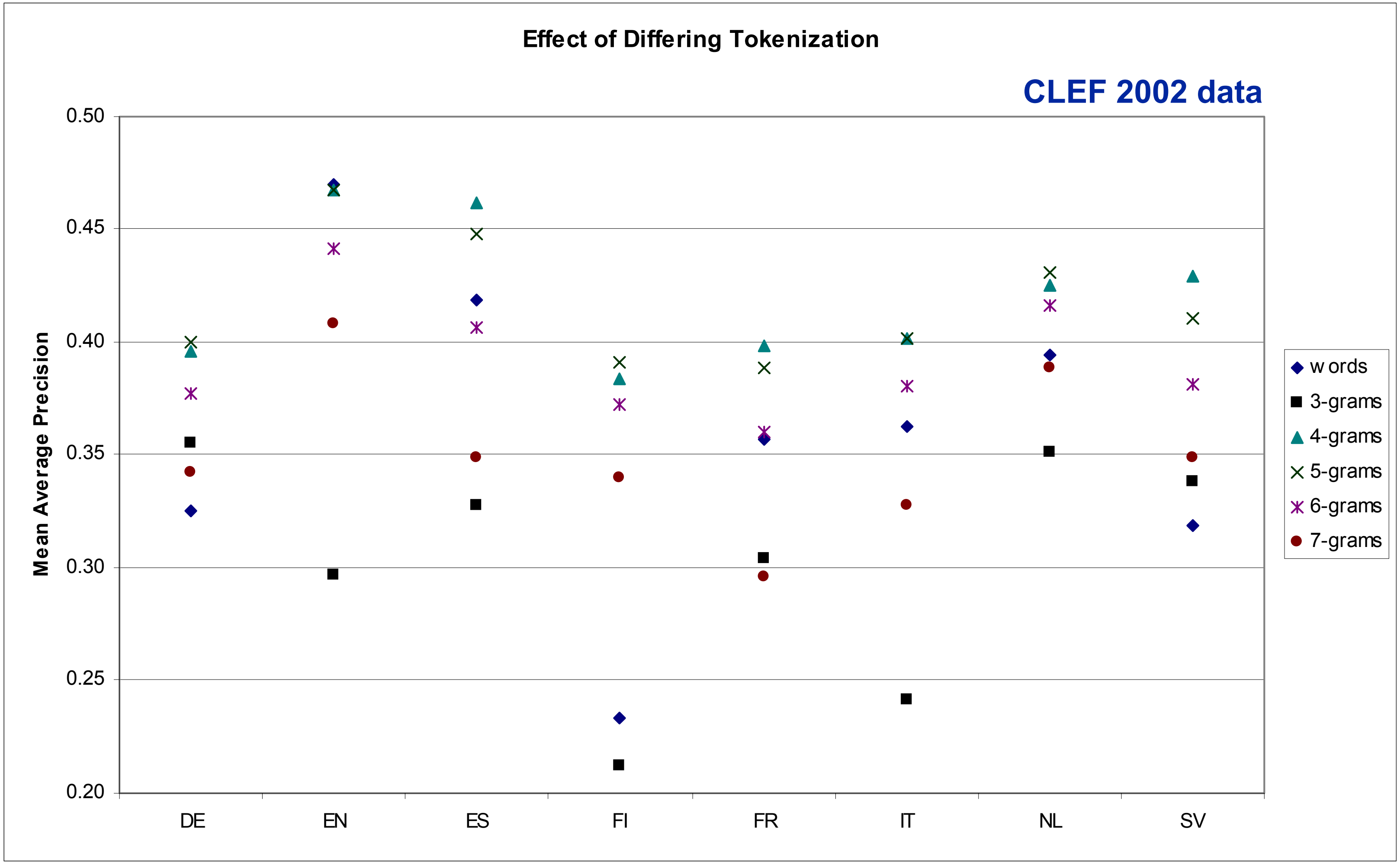
Good value for alpha: 0.5

Removal of Diacritical Marks



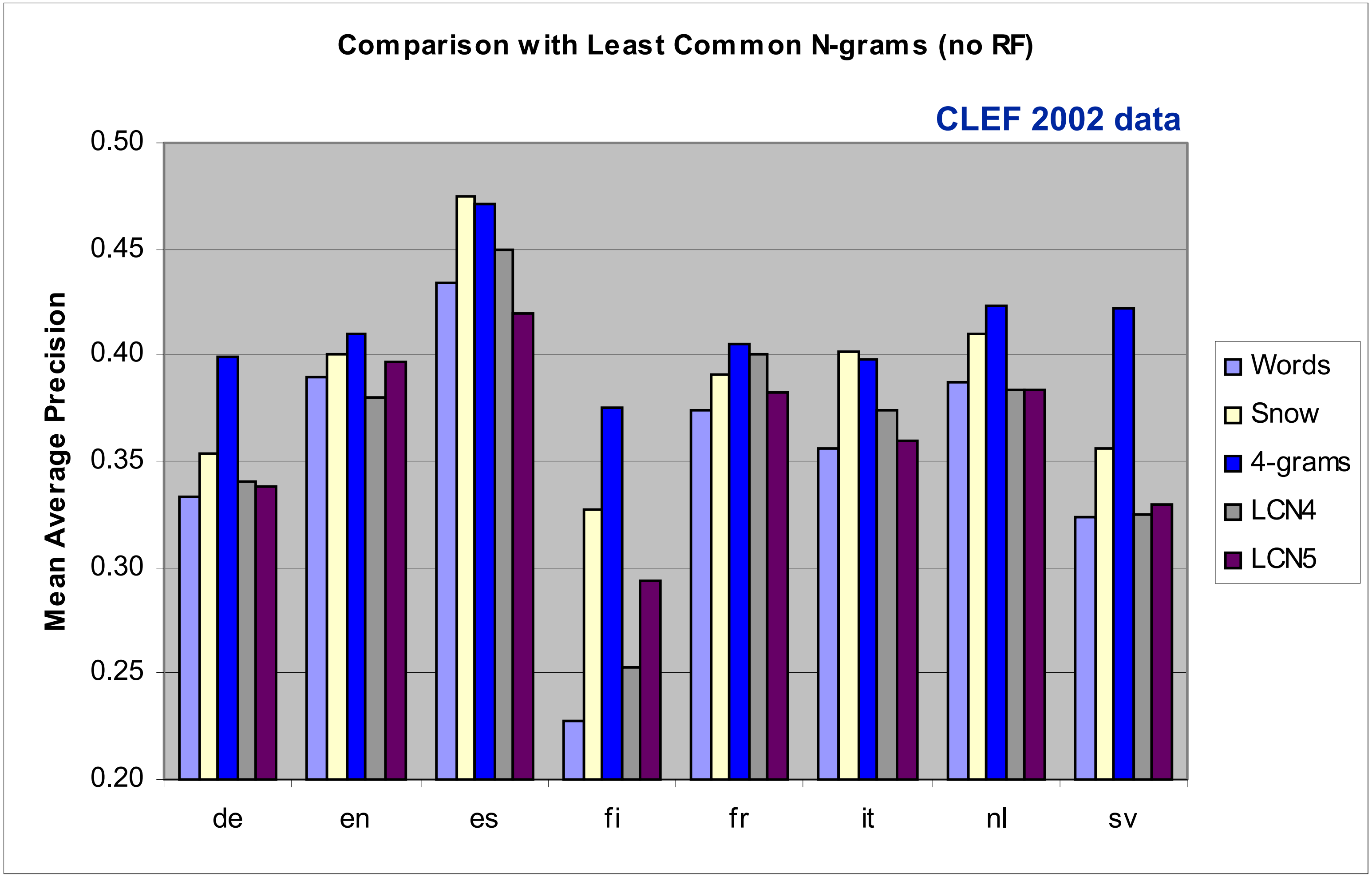
- Removal of diacritics helps in Romance languages and hurts performance in Finnish, when words are used
- Little difference is observed with 4-grams
- Tomlinson reported similar results on the CLEF 2002 data set using stems

Words vs. N-grams



For additional detail, see McNamee and Mayfield, 'Character N-gram Tokenization for European Language Text Retrieval', to appear in *Information Retrieval*.

N-grams, Stems, and Psuedo-stems



LCN4 (juggler) = 'jugg'

For additional detail, see Mayfield and McNamee, 'Single N-gram Stemming', SIGIR-03.

CLEF-2003 Monolingual Approach

- Base runs: words, stems, 4-grams, and 5-grams
  - Based on ’02 training, stems always better than words
- Submitted two runs per language
  - Runs combined using normalized scores
  - aplmoxxa: 4-grams + stems
  - aplmoxxb: 5-grams + stems
- Only *title* and *desc* fields used
- Due to a mistake in scripts, blind relevance feedback was omitted in official submissions
  - Correction and post hoc evaluation reveals general improvement with feedback

RTDC
RESEARCH & TECHNOLOGY DEVELOPMENT CENTER

CLEF 2003 Base Runs

	# topics	words	stems	4-grams	5-grams
DE	56	0.4175	0.4604	0.5056	0.4869
EN	54	0.4988	0.4679	0.4692	0.4610
ES	57	0.4773	0.5277	0.5011	0.4695
FI	45	0.3355	0.4357	0.5396	0.5498
FR	52	0.4590	0.4780	0.5244	0.4895
IT	51	0.4856	0.5053	0.4313	0.4568
NL	56	0.4615	0.4594	0.4974	0.4618
RU	28	0.2550	0.2550*	0.3276	0.3271
SV	53	0.3189	0.3698	0.4163	0.4137

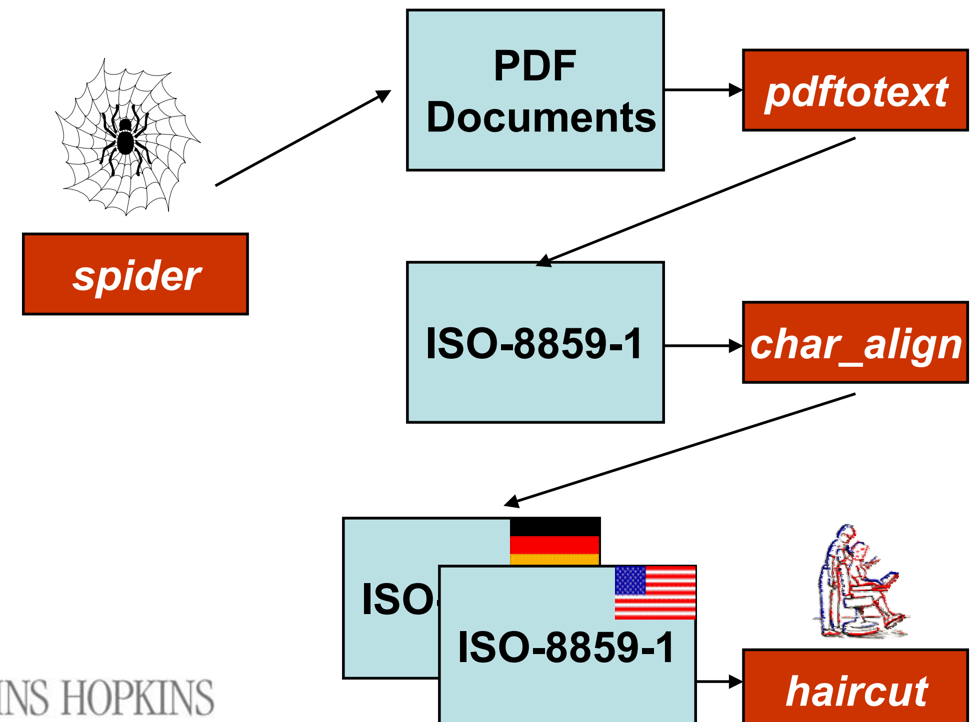
Single best monolingual technique: 4-grams

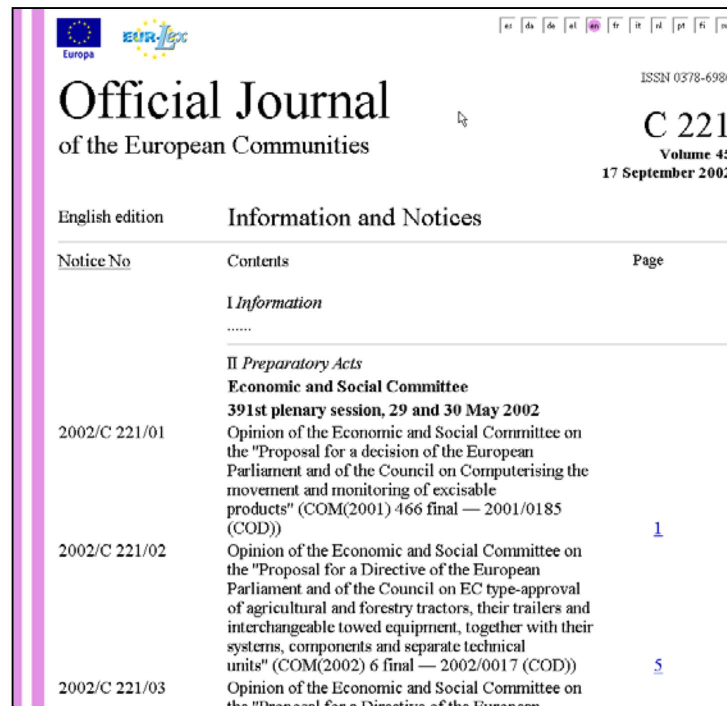
Official Monolingual Submissions

	#best	# ≥ median	MAP	Corrected (w/ RF)	%change			Best Base	Best Method
aplmodea	2	31	0.4852	0.5210	7.39%		DE	0.5056	4+s RF
aplmodeb	2	27	0.4834	0.5050	4.46%				
aplmoena	-	-	0.4943	0.5040	1.96%		EN	0.4988	5+s
aplmoenb	-	-	0.5127	0.5074	-1.03%				
aplmoesa	3	32	0.4679	0.5311	13.50%		ES	0.5277	4+s RF
aplmoesb	3	32	0.4538	0.5165	13.82%				
aplmofia	12	31	0.5514	0.5571	1.03%		FI	0.5468	5+s RF
aplmofib	9	31	0.5459	0.5649	3.49%				
aplmofra	9	35	0.5228	0.5415	3.58%		FR	0.5244	4+s RF
aplmofrb	9	37	0.5148	0.5168	0.39%				
aplmoita	7	21	0.4620	0.4784	3.54%		IT	0.5053	s RF
aplmoitb	8	22	0.4744	0.4982	5.02%				
aplmonla	3	42	0.4817	0.5088	5.63%		NL	0.4974	4+s RF
aplmonlb	2	40	0.4709	0.4841	2.86%				
aplmorua	2	17	0.3289	0.3728	10.00%		RU	0.3276	4+s RF
aplmorub	4	16	0.3282	0.3610	10.00%				
aplmosva	7	36	0.4515	0.4358	-3.47%		SV	0.4163	4+s
aplmosvb	6	38	0.4498	0.4310	-4.18%				

Building a Translation Resource

- **Mined Official Journal of EU**
  - Documents from http://europa.eu.int/
  - 33.4GB of data obtained since 12/00 (300+ MB / language)
  - Text in 11 languages produced as PDF
  - Alignments possible between any pair




Europa
EUR-Lex
ISSN 0378-6986

Official Journal
of the European Communities

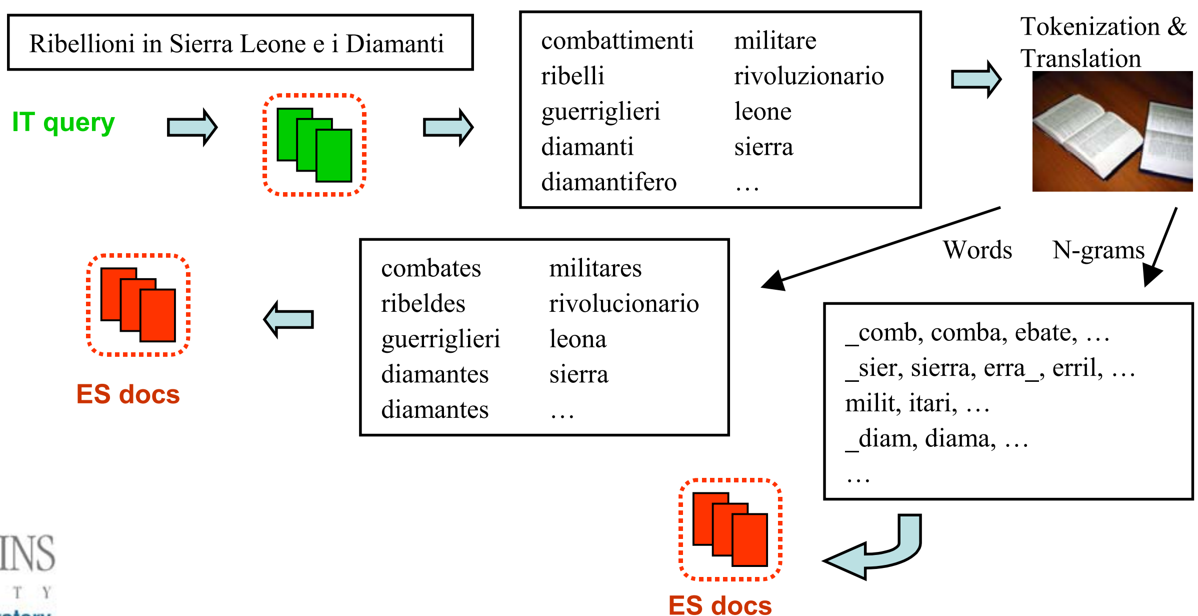
C 221
Volume 45
17 September 2002

English edition — Information and Notices

Notice No	Contents	Page
	I *Information*	
	......	
	II *Preparatory Acts*	
	Economic and Social Committee	
	391st plenary session, 29 and 30 May 2002	
2002/C 221/01	Opinion of the Economic and Social Committee on the "Proposal for a decision of the European Parliament and of the Council on Computerising the movement and monitoring of excisable products" (COM(2001) 466 final — 2001/0185 (COD))	1
2002/C 221/02	Opinion of the Economic and Social Committee on the "Proposal for a Directive of the European Parliament and of the Council on EC type-approval of agricultural and forestry tractors, their trailers and interchangeable towed equipment, together with their systems, components and separate technical units" (COM(2002) 6 final — 2002/0017 (COD))	5
2002/C 221/03	Opinion of the Economic and Social Committee on the "Proposal for a Directive of the European	

CLEF-2003 Bilingual Approach

- **Bilingual Task**
  - **Pre-translation expansion performed using source language subcollection; words extracted**
  - **Words tokenized and tokens translated (1-best)**
  - **Used only aligned corpus for *direct* translation**
  - **Formed hybrid runs by merging techniques**

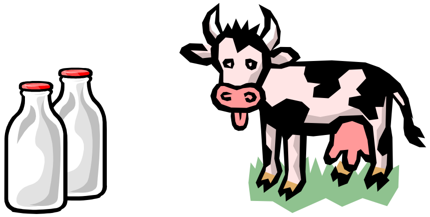


N-gram Translations

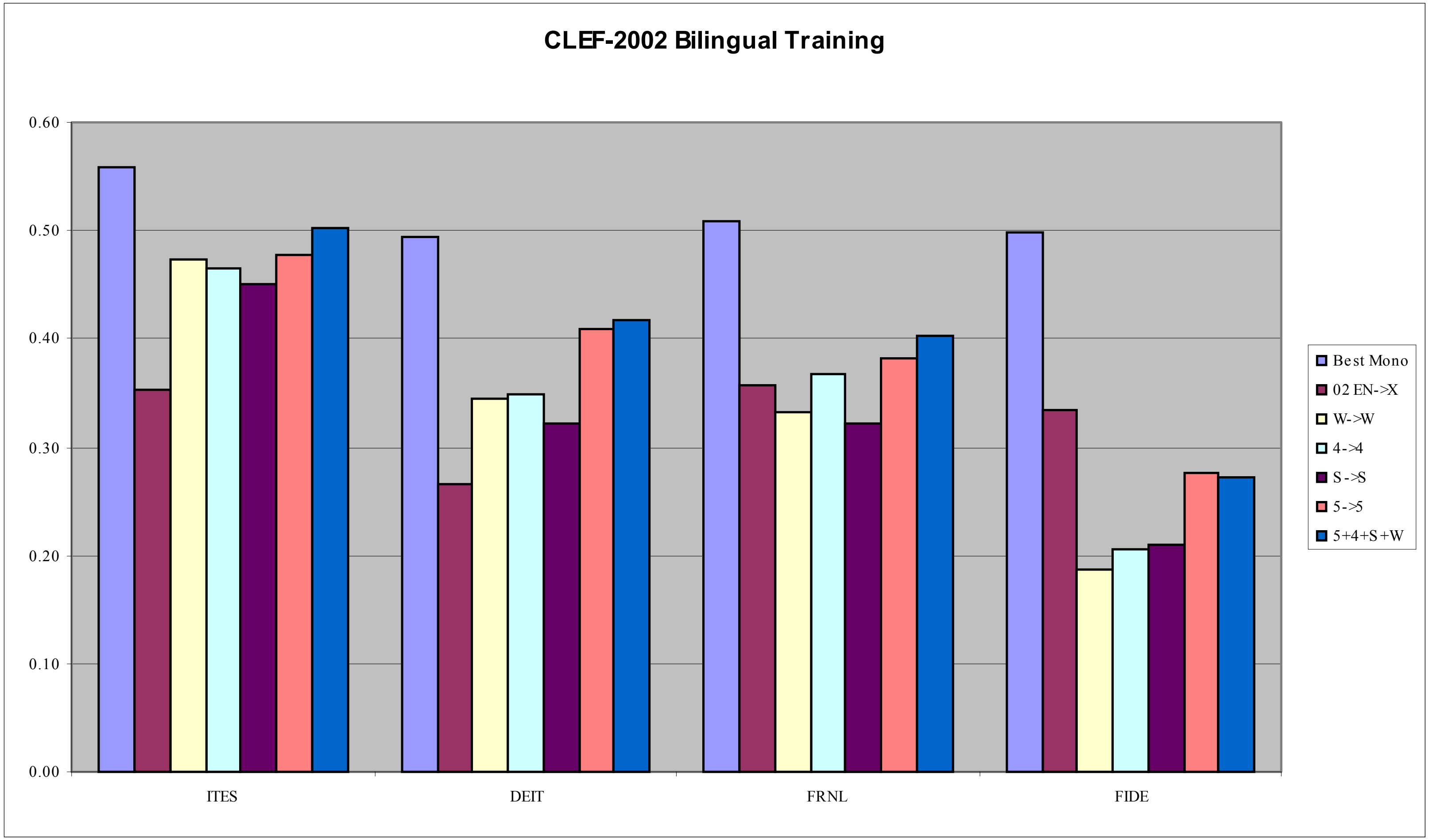
- Just as words can be statistically translated using an aligned bitext, so can n-grams
- Difficult to quantify accuracy of mappings
- May mitigate problems in dictionary-based CLIR
  - word lemmatization
  - multiword expressions
  - out of vocabulary words, particularly names

	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_



Token-specific translation



Bilingual Submissions

	tokens	RF	#best	# ≥ median	#topics	MAP	% mono
aplbideita	w+s+4+5	Yes	11	38	51	0.4264	89.88%
aplbideitb	w+s+4+5	No	12	45	51	0.4603	97.03%
aplbifidea	w+s+4+5	Yes	16	39	56	0.3454	71.19%
aplbifideb	w+s+4+5	No	16	42	56	0.3430	70.69%
aplbifrnla	w+s+4+5	Yes	15	33	56	0.4045	83.97%
aplbifrnlb	w+s+4+5	No	13	33	56	0.4365	90.62%
aplbiitesa	w+s+4+5	Yes	5	32	57	0.4242	90.66%
aplbiitesb	w+s+4+5	No	4	38	57	0.4261	91.07%

Source language queries were expanded to 60 words using the appropriate sub-collection. Words were then optionally tokenized, and each token was translated directly to a corresponding token in the target language. Target language retrieval was then performed, and additional post translation relevance feedback was optionally applied. Finally the runs corresponding to the four term types were merged.

Multilingual Retrieval

- We applied the same general methods used on the bilingual task
  - English was used as the source language
  - Only 4-grams, words, and stems were used as base runs.
    - We ran out of time building 5-gram translations for the eight languages
    - Probably lowered our performance
- A hybrid run was constructed for each target language
- These four (eight) runs were then merged by re-normalized scores.

RTDC
RESEARCH & TECHNOLOGY DEVELOPMENT CENTER

"A Basic Novice Solution"

CLEF

WHAT'S NEXT

...m Uzbek to Klingon, the Machine Cracks the Code

...R JOHN FARAH

...99, at a workshop on ... translation at Johns ...ty, Kevin Knight ...an advertisement to ...search team he was ...f the ad was a picture ...archment covered in ... "To most people, this ...," the ad announced. ...e broken." ...duct yet to be created ...our in a new bunch of ..., alongside a picture ...e think you'll be sur-...

...eant to be a motiva-... the field of statistical ...s all but dead. In the ...ssed since that work-...ad of machine trans-... University of South-...ation Sciences Insti-... how prophetic the ad ...are," he said. "It's no ...

... translation — in ...tially learn new lan-...n instead of being ... by bilingual human ... taken off. The new ...tists to develop ma-...ms for a wide num-...es at a pace that ex-...possible. ...rs said the progress ...ical machine transla-...ssed that of the tradi-...

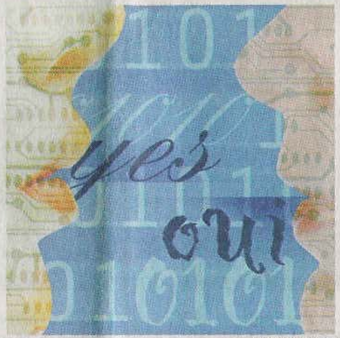
tional machine translation programs used by Web sites like Yahoo and BabelFish. In the past, such programs were able to compile extensive databanks of foreign languages that allowed them to outperform statistics-based systems.

Traditional machine translation relies on painstaking efforts by bilingual programmers to enter the vast wealth of information on vocabulary and syntax that the computer needs to translate one language into another. But in the early 1990's, a team of researchers at I.B.M. devised another way to do things: feeding a computer an English text and its translation in a different language. The computer then uses statistical analysis to "learn" the second language.

Compare two simple phrases in Arabic: "rajl kabir" and "rajl tawil." If a computer knows that the first phrase means "big man," and the second means "tall man," the machine can compare the two and deduce that rajl means "man," while kabir and tawil mean "big" and "tall," respectively. Phrases like these, called "N-grams" (with N representing the number of terms in a given phrase) are the basic building blocks of statistical machine translation.

Although in one sense it was more economical, this kind of machine translation was also much more complex, requiring powerful computers and software that did not exist for most of the 90's. The Johns Hopkins workshop changed all that, yielding a software application package, Egypt/Giza, that made statistical translation accessible to researchers across the country.

"We wanted to jump-start a vibrant field," Dr. Knight said. "There was no software or data to play with."



Mary Ann Smith

Today researchers are racing to improve the quality and accuracy of the translations. The final translations generally give an average reader a solid understanding of the original meaning but are far from grammatically correct. While not perfect, statistics-based technology is also allowing scientists to crack scores of languages in a fraction of the time, and at a fraction of the cost, that traditional methods involved.

A team of computer scientists at Johns Hopkins led by David Yarowsky is developing machine translations of such languages as Uzbek, Bengali, Nepali ... "Star Trek."

"If we can learn how t... Klingon into English, then ... guages are easy by comp... "All our techniques require ... two languages. For exam...

Language Institute translated 'Hamlet' and the Bible into Klingon, and our programs can automatically learn a basic Klingon-English MT system from that."

Dr. Yarowsky said he hoped to have working translation systems for as many as 100 languages within five years. Although the grammatical structures of languages like Chinese and Arabic make them hard to analyze statistically, he said, it will only be a matter of time before such hurdles are overcome. "At some point, we start encountering the same problems over and over," he said.

In addition to the release of Egypt/Giza in

Armed with an English text and a translation, a computer uses statistical analysis to 'learn' the second tongue.

1999, the spread of the Internet has led to an explosion of translated texts in far-flung languages, greatly aiding the team's research. Researchers have also benefited from a much faster means of evaluating the outcome of translation experiments: a computerized technique developed by I.B.M. enables researchers to test 10 to 100 new approaches for cracking languages each day.

provides scientists with a fast, objective measurement that they can use to note improvement and saves them from having to review every unsuccessful experiment.

"Before Bleu, it was really a bad state of affairs," said Alex Fraser, a doctoral student at U.S.C. "You look at broken couplets of English for a long time, and eventually you start to accept it more and more."

Despite the progress being made in statistical machine translation, some researchers remain skeptical, preferring to focus their efforts on language-specific translation techniques. Ophir Frieder, a professor of computer science at the Illinois Institute of Technology, is working on a search system exclusive to Arabic text.

"Yes, N-grams work on any language, but as a search technique they work poorly on every language," he said. "It's a basic novice solution."

Dr. Knight acknowledges that statistical machine translation is far from perfect. In its latest efforts, his team has sought to combine the statistical and traditional approaches to achieve ...imum accuracy and to produce trans... that the average computer user can ...stand. The best machine translation ...ms today, while capable of yield... ...age's general meaning, are better ...n for their muddled syntax than their accuracy. By applying the principles of statistical translation to

"Yes, N-grams work on any language, but as a search technique they work poorly on every language," he said. "It's a basic novice solution."

-quote attributed to an IR researcher in the New York Times on 31 July 2003

JOHNS HOPKINS
UNIVERSITY
Applied Physics Laboratory

Conclusions

- **When retrieval accuracy is of greatest import, n-grams are recommended for monolingual tasks**
  - **Generally outperform plain words and Snowball-produced stems**
  - **N=4 or N=5 both highly effective across CLEF languages**
- **Bilingual retrieval with n-grams is also attractive**
  - **5-gram translation alone does very well**
  - **Avoids problems specific to word-based retrieval**
- **Computational issues should be addressed**