Resolving Translation Ambiguity using Monolingual Corpora

Yan Qu, Gregory Grefenstette, David A. Evans

Clairvoyance Corporation

September 19, 2002

Overview of Participation

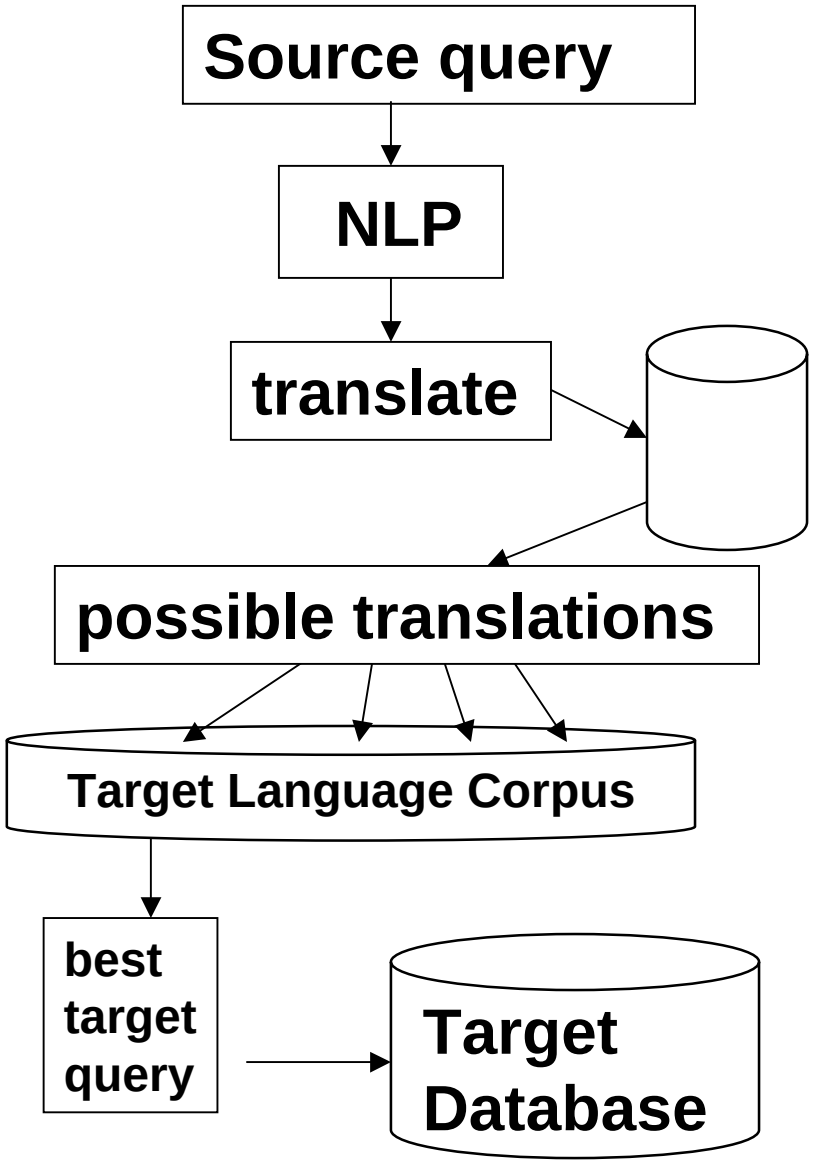
- **First-time participation in CLEF**
- **Spanish-to-English**

  Chinese-to-English
  - automatic runs
- **Research focus:**

  Evaluating methods for selecting best translations from translation glosses

Our Approach to CLIR

- **NLP (Pseudo-NLP)**
  - Spanish
    - Tokenizer
    - Morphological analysis
  - Chinese
    - Word segmentation
- **Query translation**
- **Bilingual dictionary**
- **Monolingual target corpora**
- **How to choose translations**



Clarit English NLP

- **Used for processing the English corpus**
- **Consists of a parser and morphological analyzer**
- **Uses an English lexicon and grammar to identify linguistic structures in texts**
- **Supplemented by a "stop word" list to filter out substantive words that are extraneous to the topics (e.g., *document*, *relevant*)**
  - Pre-CLEF2002 stop word list

Spanish Topic Processing

Topic 136: ***Torre inclinada de Pisa. ¿En qué estado se encuentra la torre inclinada de Pisa?***

1. **Tokenized and morphologically analyzed: keeping nouns, verbs, adjectives and adverbs**
2. **Remove stop words**

   torre, inclinar, pisa, estado, torre, inclinar, pisa.

3. **Get translations by dictionary lookup**

Dictionary roughly compiled from Internet sources

estado	inclinar	pisa	torre
state states statis	apt, bow, drooping, incline, inclined, inclining, sloping, stooping, titling, verging	pisa	high tower towers

Chinese Topic Processing

Topic 136: 比萨斜塔; 比萨斜塔的健康情况如何？

1. Word segmentation – dictionary-based, greedy algorithm

比萨；斜；塔；比萨；斜；塔；的；健康；情况；如何；？；

2. Stop words removal

比萨；斜；塔；比萨；斜；塔；健康；情况；

3. Get translations by dictionary lookup

Dictionary obtained from LDC:
www.ldc.upenn.edu/Projects/Chinese/LDC_ch.htm#e2cdict

比萨	斜	塔	健康	情况
pisa	**askant** **slanting**	**ter** **pagoda** **tower**	**hygenia** **exuberance** **health** **healthiness**	**circumstance** **circumstantiality** **situation** **state** **affairs**

Translation Disambiguation

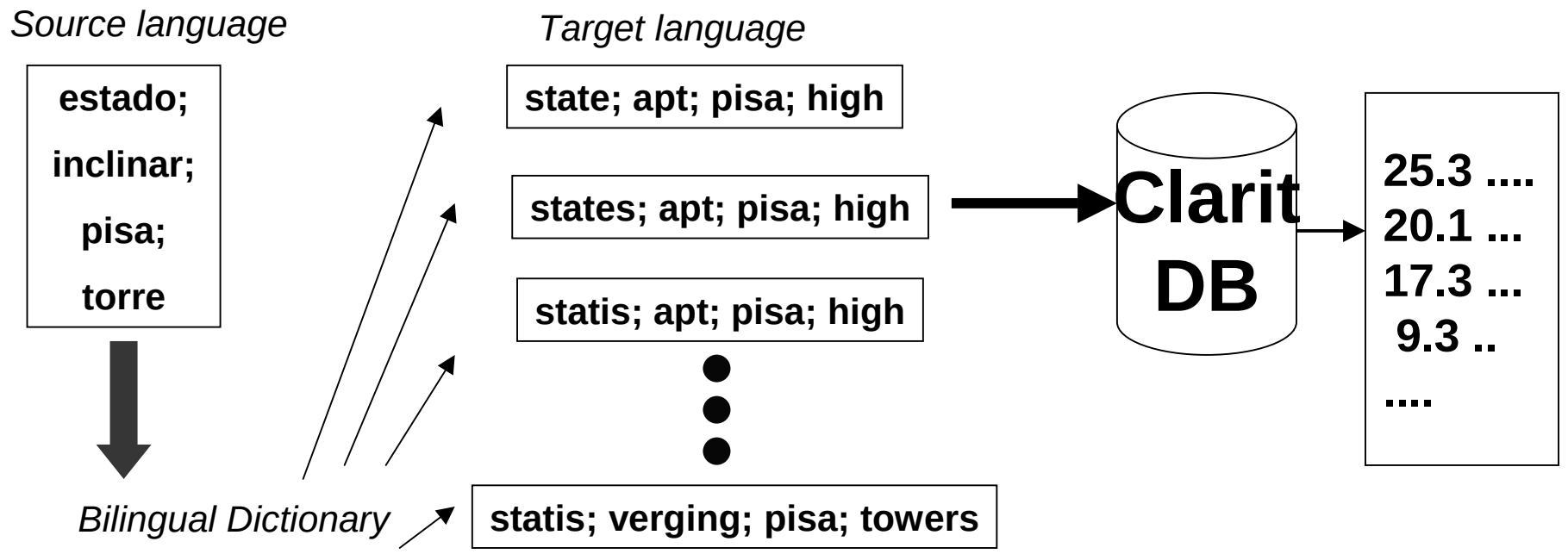
Two Approaches

- **Use Local Database Scoring, Decide Best Translations**
  - Evans, David A., 2000; 2001
- **Use WWW to Decide**
  - Grefenstette, Gregory, 1999

The Approaches

- **Translate query words**
- **Test combinations against target language database (Clarit/WWW)**
- **Select the combination with the best "*coherence score*", based on**
  - Web page counts (Web)
  - Retrieval scores (Corpus1)
  - Mutual information scores (Corpus2)

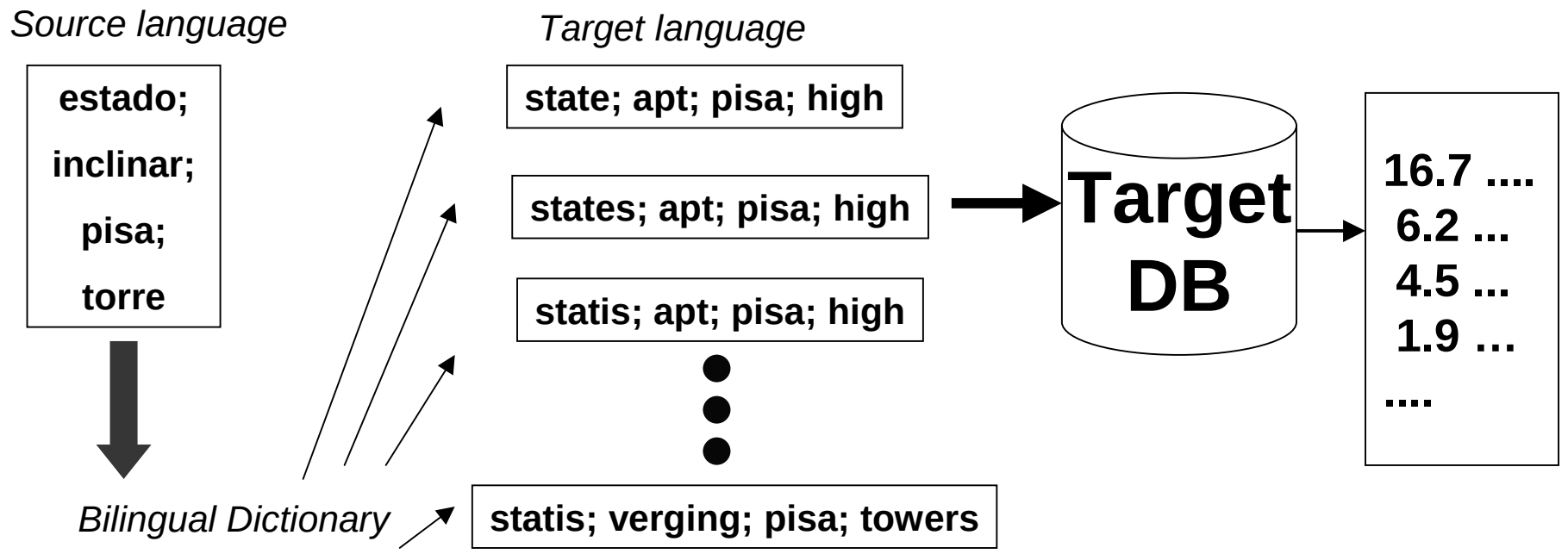
The Corpus1 Approach



estado	inclinar	pisa	torre
state	apt	pisa	high
states	bow		tower
statis	incline		towers
	....		
	verging		

Without knowing which documents are relevant, take queries that provide the richest retrieval scores

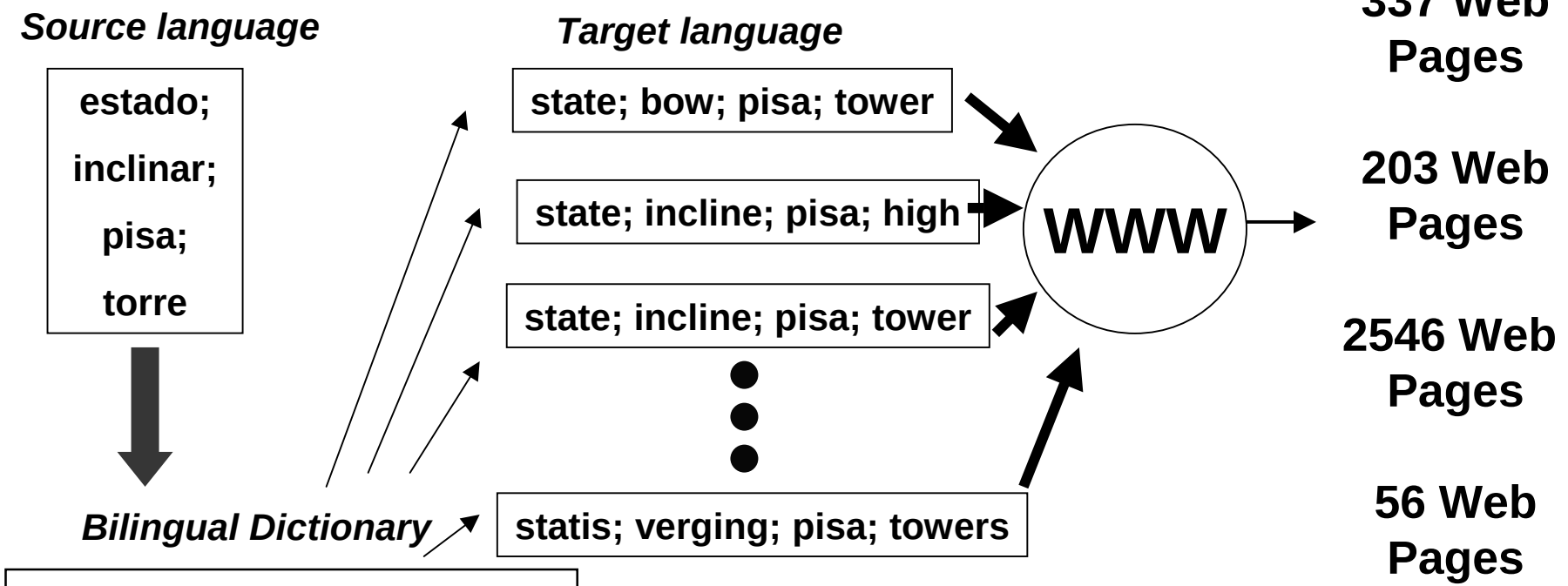
The Corpus2 Approach



estado	inclinar	pisa	torre
state	apt	pisa	high
states	bow		tower
statis	incline		towers
	....		
	verging		

Select sequences with the highest sum of pairwise mutual information scores between terms

The Web Approach



estado	inclinar	pisa	torre
state	apt	pisa	high
states	bow		tower
statis	incline		towers
	....		
	verging		

Take queries in which words appear together most often on the WWW

Bells and Whistles

- **Using full morphology on the web**
- **Trade – trade, trades, traded, trading**

81 China_USA AND (trading OR traded OR trade OR trades) AND (balancing OR balance OR balanced OR balances) AND (issued OR issues OR issuing OR issue)

71 China_USA AND (trading OR traded OR trade OR trades) AND (balancing OR balance OR balanced OR balances) AND (problems OR problem)

24 China_USA AND (trading OR traded OR trade OR trades) AND (balancing OR balance OR balanced OR balances) AND (topic OR topics)

5 China_USA AND (trading OR traded OR trade OR trades) AND (equilibriums OR equilibrium OR equilibria) AND (issued OR issues OR issuing OR issue)

4 China_USA AND (trading OR traded OR trade OR trades) AND (equilibriums OR equilibrium OR equilibria) AND (problems OR problem)

3 China_USA AND (trading OR traded OR trade OR trades) AND (equilibriums OR equilibrium OR equilibria) AND (topic OR topics)

Dealing with Long Queries

Original Terms	S1	S2	S3	S4	S5
Target Language translations	T11 T12 T13	T21 T22	T31 T32 T33 T34	T41	T51 T52

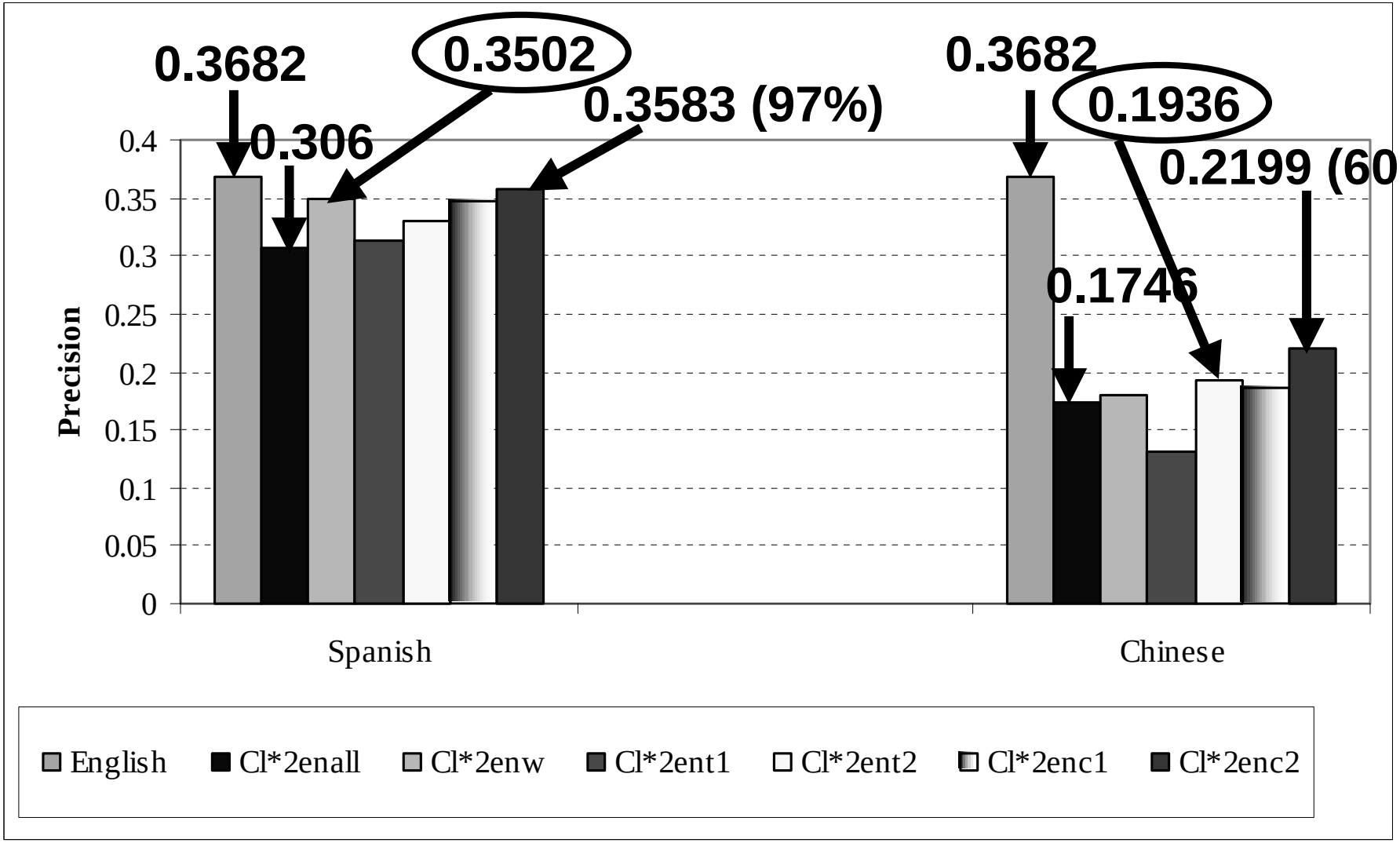
Score subsequences, rather than all combinations

T11 NEAR T21 NEAR T31
T12 NEAR T21 NEAR T31
T13 NEAR T21 NEAR T31
T11 NEAR T22 NEAR T31
T12 NEAR T22 NEAR T31
T13 NEAR T22 NEAR T31
T21 NEAR T31 NEAR T41 score : 35
T22 NEAR T31 NEAR T41 score : 4456
T21 NEAR T32 NEAR T41
T22 NEAR T32 NEAR T41
...

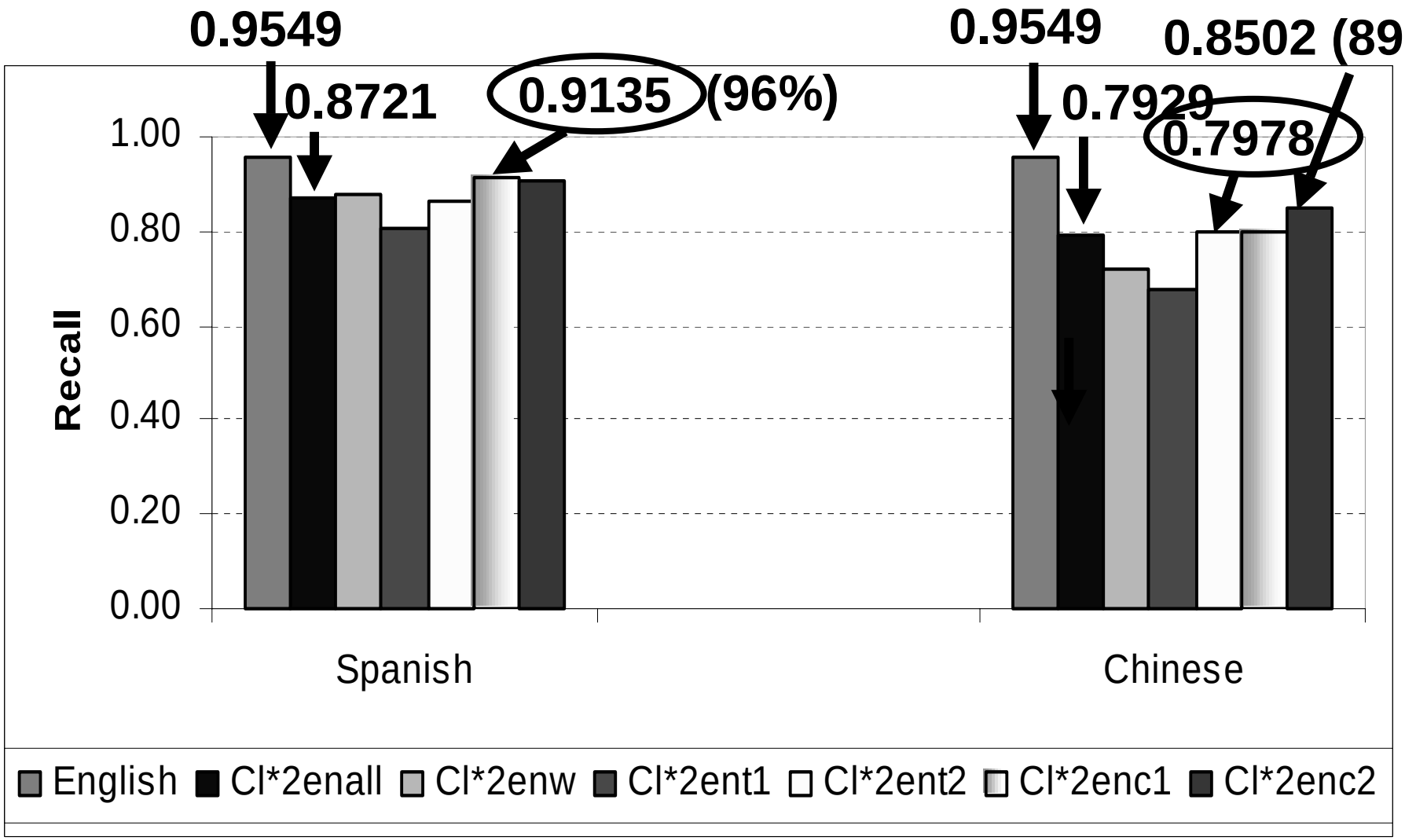
Experiment Runs

- **English –** English monolingual baseline
- **Cl*2enall –** Spanish/Chinese-to-English bilingual retrieval baseline with all possible translations
- **Cl*2enw –** Spanish/Chinese-to-English bilingual retrieval using the Web method
- **Cl*2ent1 –** Spanish/Chinese-to-English bilingual retrieval using the Corpus1 method
- **Cl*2ent2 –** Spanish/Chinese-to-English bilingual retrieval using the Corpus2 method
- **Cl*2enc1 –** Best translations from the Web method and the Corpus1 method
- **Cl*2enc2 –** Best translations from the Web method, the Corpus1 method, and the Corpus2 method
- **All runs are automatic, with pseudo relevance feedback, using Title+Description fields.**

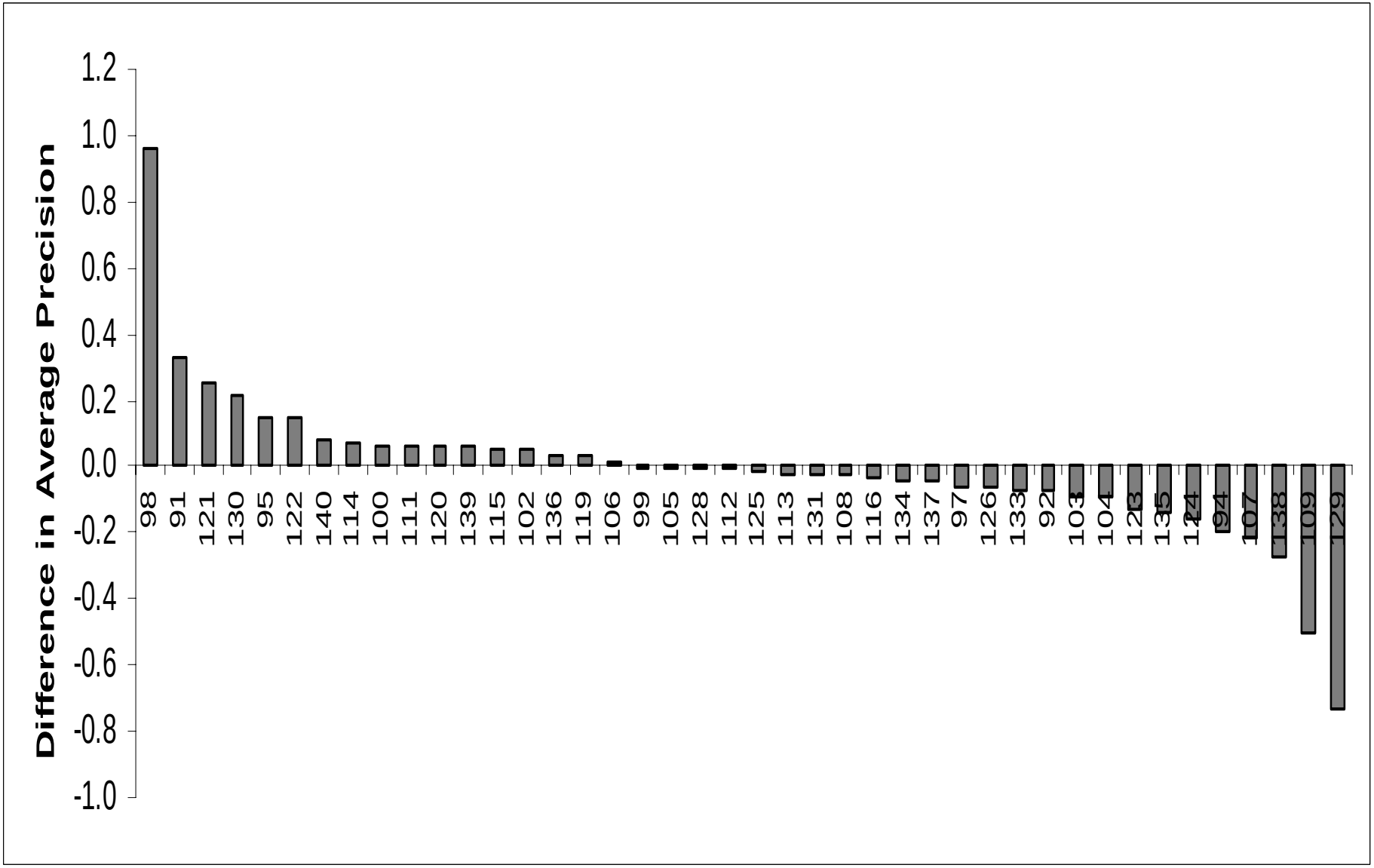
Comparison – Average Precision



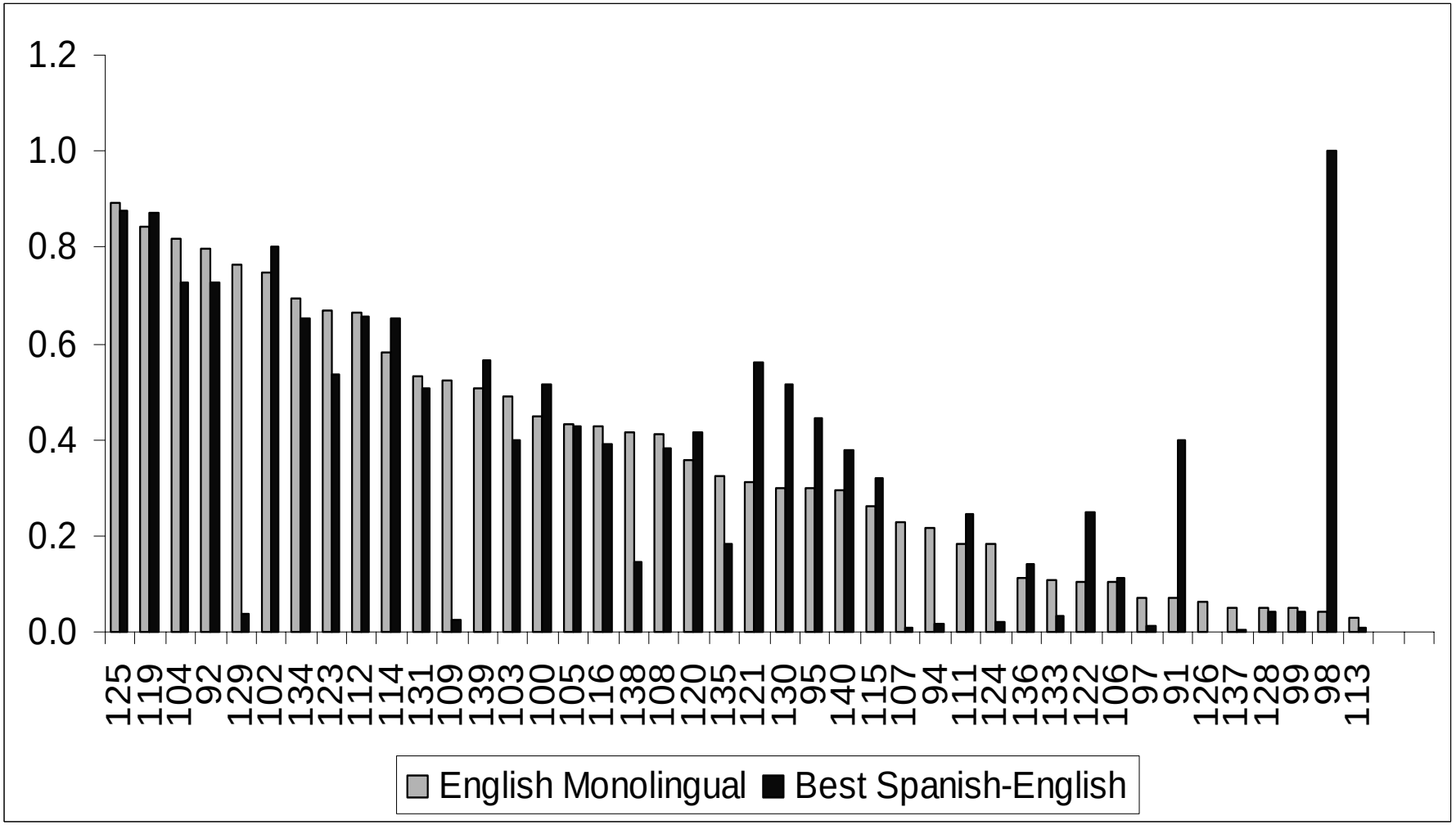
Comparison – Recall



Spanish-English vs English Monolingual



Query-by-Query Precision



Spanish-English: Better Cases

- **Different word choice**

  ***Population* in the English version of Topic 95 (Conflict in Palestine) becomes *town* in the Spanish-to-English translation of *poblacion*.**

  ***Ski races* in the English version of Topic 102 becomes *ski competition* in the translations.**

- **Different word ordering**

  ***Grunge rock* appears in the translation of Topic 130 while the English contains *grunge group*.**

- **Different formulations of topics**

  Topic 138 (Foreign words in French) contains *lengua*, which translates to *language* not present in the English version.

- **Different Proper Names**

  – **Topic 98 Spanish: *Kaurismaki* English: *Kaurismäki***

Spanish-English: Error Analysis

- **Proper names written differently and missing in dictionary**

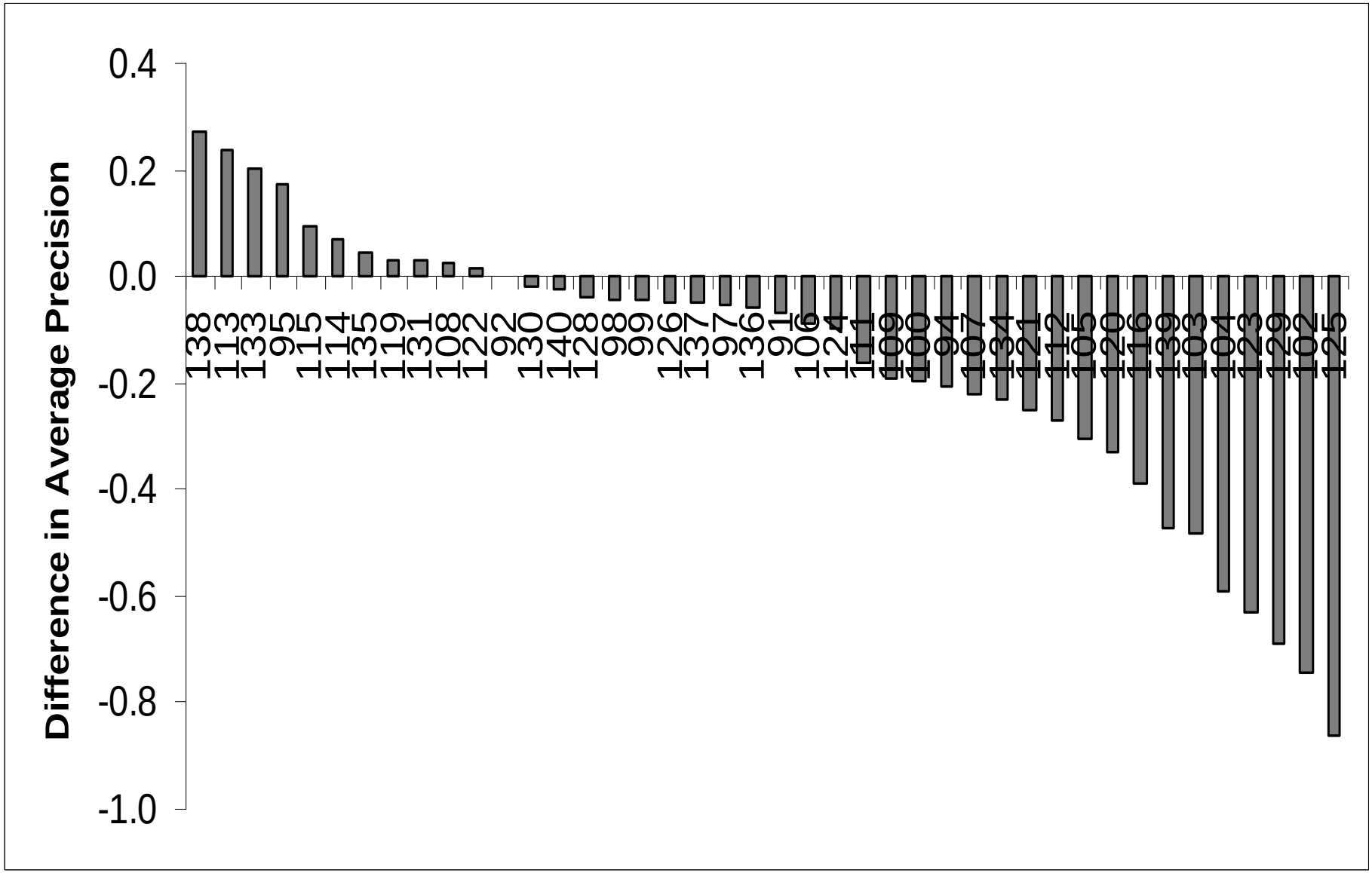
  Topic 94 (Solzhenitsyn):
  Spanish: *Solzhenitsin*
  English: *Solzhenitsyn*

  Topic 113 (European Cup):
  Spanish: *Eurocopa*
  English: *European Cup*

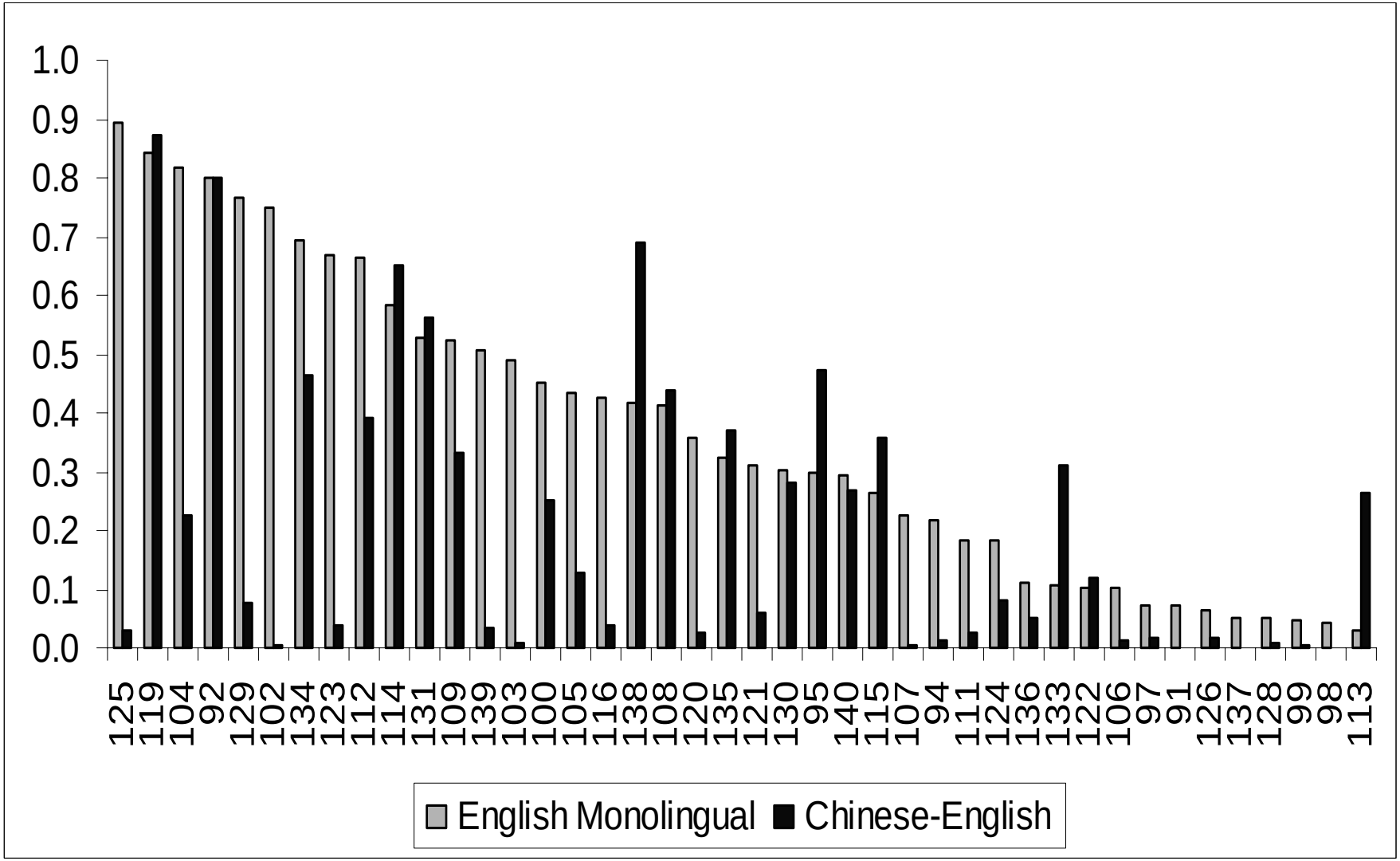
- **Dictionary divergences**

  Topic 107 (Genetic Engineering):
  alimentaria* missing translation *food
  while *cadena alimentaria* is translated as *food chain*

Chinese-English vs English Monolingual



Query-by-Query Precision



Chinese-English: Better Cases

- **Reduced ambiguity in translation**

  Topic 113 (European Cup):

  English: *football*

  Chinese: *soccer*

- **Difference in word choice**

  Topic 133 (German Armed Forces Out-of-area):

  English: *area* (out-of-area)

  Chinese: *border*

Chinese-English: Error Analysis

- **Improper segmentation of words**

  Topic 111: 动画 **(animation)**
  -- 动 **(act, arouse, move, etc)**
  -- 画 **(draw, painting, picture)**

  Topic 129: 银河 **(galaxy)**
  -- 银 **(silver)**
  -- 河 **(river)**

- **Improper segmentation of transliterated names**

  Topic 102: 阿尔伯特·托姆巴

  -- 阿 ; 尔 ; 伯 ; 特 ; · ; 托 ; 姆 ; 巴

- **Different word choice**

  Topic 107: 遗传 **(genetic)**
  -- transmit, transmittal, hereditary
  Topic 99: 屠杀 **(Holocaust)**
  -- butchery, massacre

Summary

- **We explored methods for selecting best translations that take advantage of co-occurrence statistics.**
- **Relatively simple and available resources (e.g., Web, target corpus) can produce remarkably high CLIR performance.**

  Spanish-English

  Avg. Precision = 0.3583, 97% of English monolingual

  Recall = 0.9062, 95% of English monolingual

  Chinese-English

  Avg. Precision = 0.2199, 60% of English monolingual

  Recall = 0.8502, 89% of English monolingual
- **Quality of the translation resources has a big impact on the performance of CLIR.**

Future Work

- **Identify the optimal context span**
- **Find ways to preliminarily prune translations and paths**
- **Identify optimal retrieval cutoff point for the corpus-based method**
- **Incorporate automatic phrasal translations**
- **Use better segmentation methods and identify names in Chinese processing**

The End