Multilingual Information Retrieval Using English and Chinese Queries

Aitao Chen

School of Information Management and Systems

University of California, Berkeley

CLEF 2001 Workshop: 3-4 Sept, 2001, Darmstadt, Germany

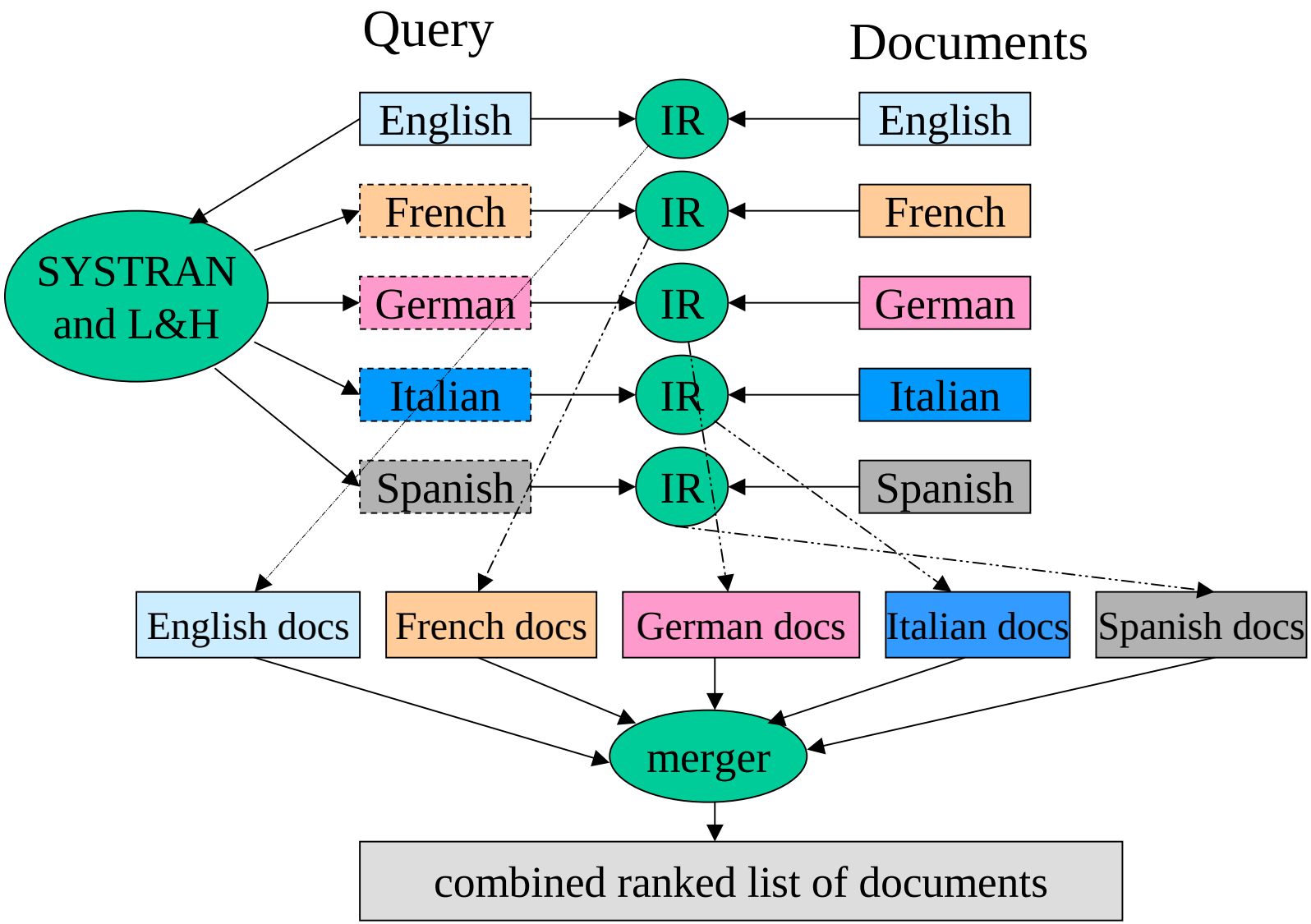
Outline

- Overview over what we did at CLEF-2001
- German decompounding
- Chinese topics translation
- Merging strategies and alternative methods
- Conclusions

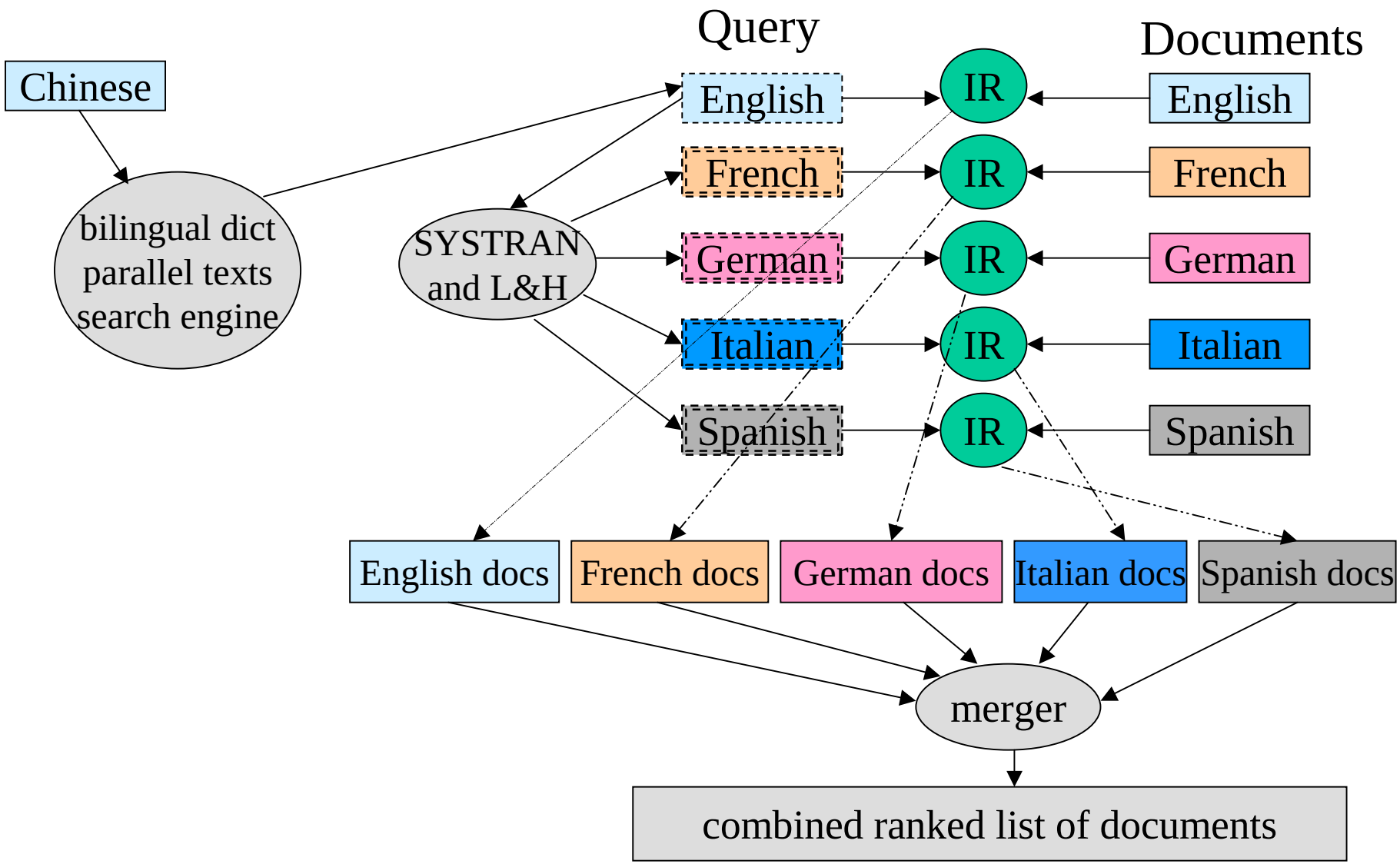
Participation in CLEF-2001

- Monolingual task (German and Spanish)
- Bilingual task (Chinese to English)
- Multilingual task (English and Chinese)

Overview of Multilingual Information Retrieval Using English Queries



Overview of Multilingual Information Retrieval Using Chinese Queries



German Decompounding Procedure

- Create a German base dictionary consisting of single words only (compounds are excluded).
- Decompose a compound into component words found in the German base dictionary.
- Choose the decomposition with the minimum number of component words.
- If there are more than one decompositions having the minimum number of component words, choose the decomposition with the highest probability.

German Decompounding: Example 1

Compound: filmfestspiele (film festival)

1. Base dictionary

…
film
fest
fests
festspiele
piele
s
…

2. Decompositions:

1. film fest s piele
2. film fest spiele
3. film fests piele
4. film festspiele

3. Result:

filmfestspiele = file festspiele

German Decompounding: Example 2

Compound: hungerstreiks (hunger strike)

1. Base dictionary

erst
hung
hunger
hungers
hungerst
reik
reiks
s
streik
streiks

2. Decompositions: log p(D)

1.	hung erst reik s	-55.2
2.	hung erst reiks	-38.0
3.	hunger streik s	-38.7
4.	hunger streiks	-21.4
5.	hungerst reik s	-52.1
6.	hungerst reiks	-34.9

3. Result:

hungerstreiks = hunger streiks

German Decompounding: Probability of Decomposition

$$C = W_1 \; W_2 \; W_3 \; W_4$$

$$p(C) = p(W_1) * p(W_2) * p(W_3) * p(W_4)$$

$$p(w) = \frac{tfc(w)}{\sum_{i=1}^{n} tfc(w_i)}$$

tfc(w) is the number of times word w occurs in a corpus.

n is the number of unique words (including compounds) in a corpus.

German Decompounding: Failed Cases

1. erdatmosphäre = erde + atmosphäre

(earth atmosphere)

2. mittagessenzeit = mittag essen zeit
(noon meal time)

(mittagessenzeit = mittagessen zeit)
lunch time

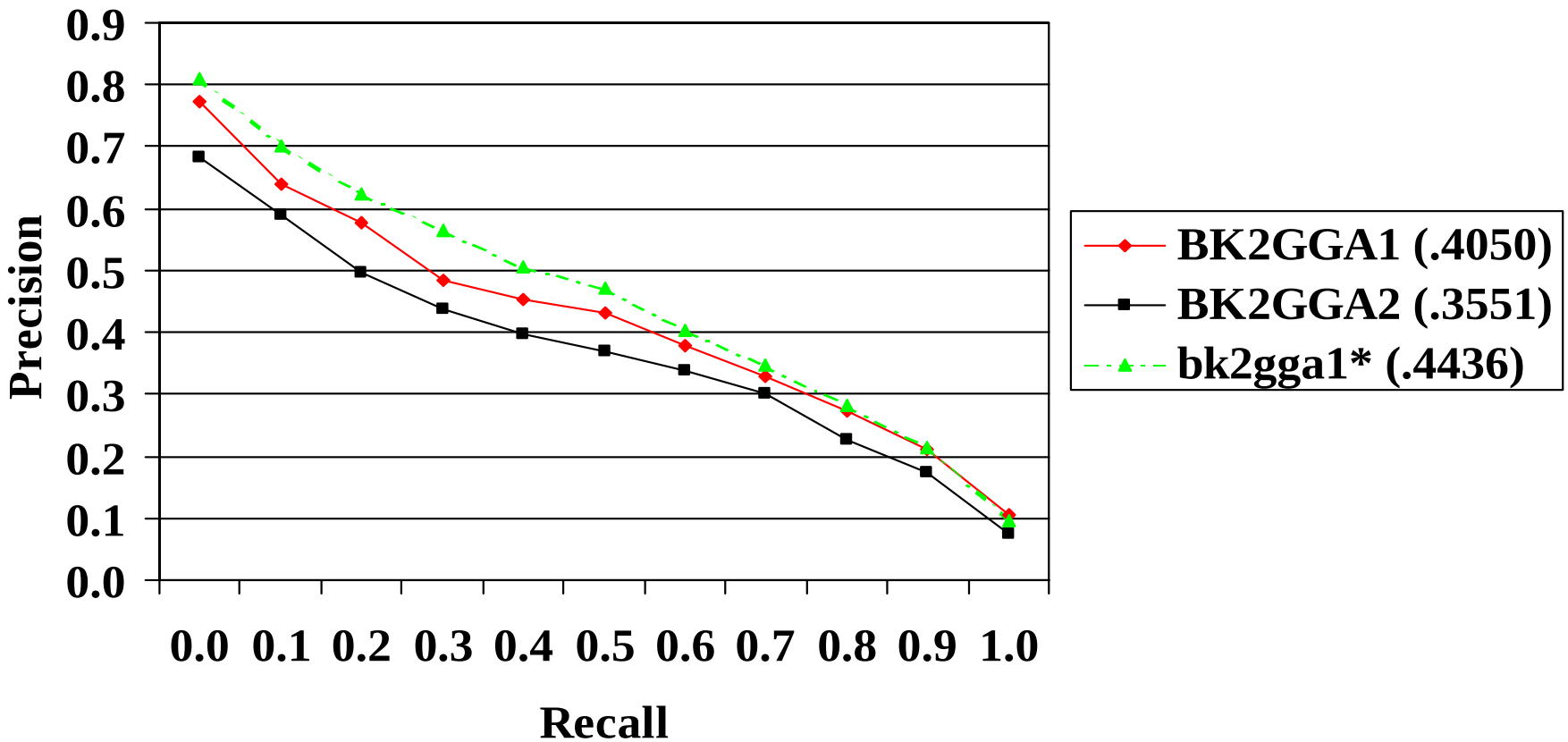
3. And others

German Decompounding and Monolingual Retrieval Performance

Test collections	-Decompounding -Stemming -Expansion	+ Decompounding	Change
CLEF-2001 (49/225K)	.3673 (1877/2130)	.4314 (1949/2130)	+17.45%
CLEF-2000 (37/154K)	.3189 (673/821)	.4112 (770/821)	+28.94%
TREC-6/7/8 (73/252K)	.2993 (1907/2626)	.3368 (2172/2626)	+12.53%

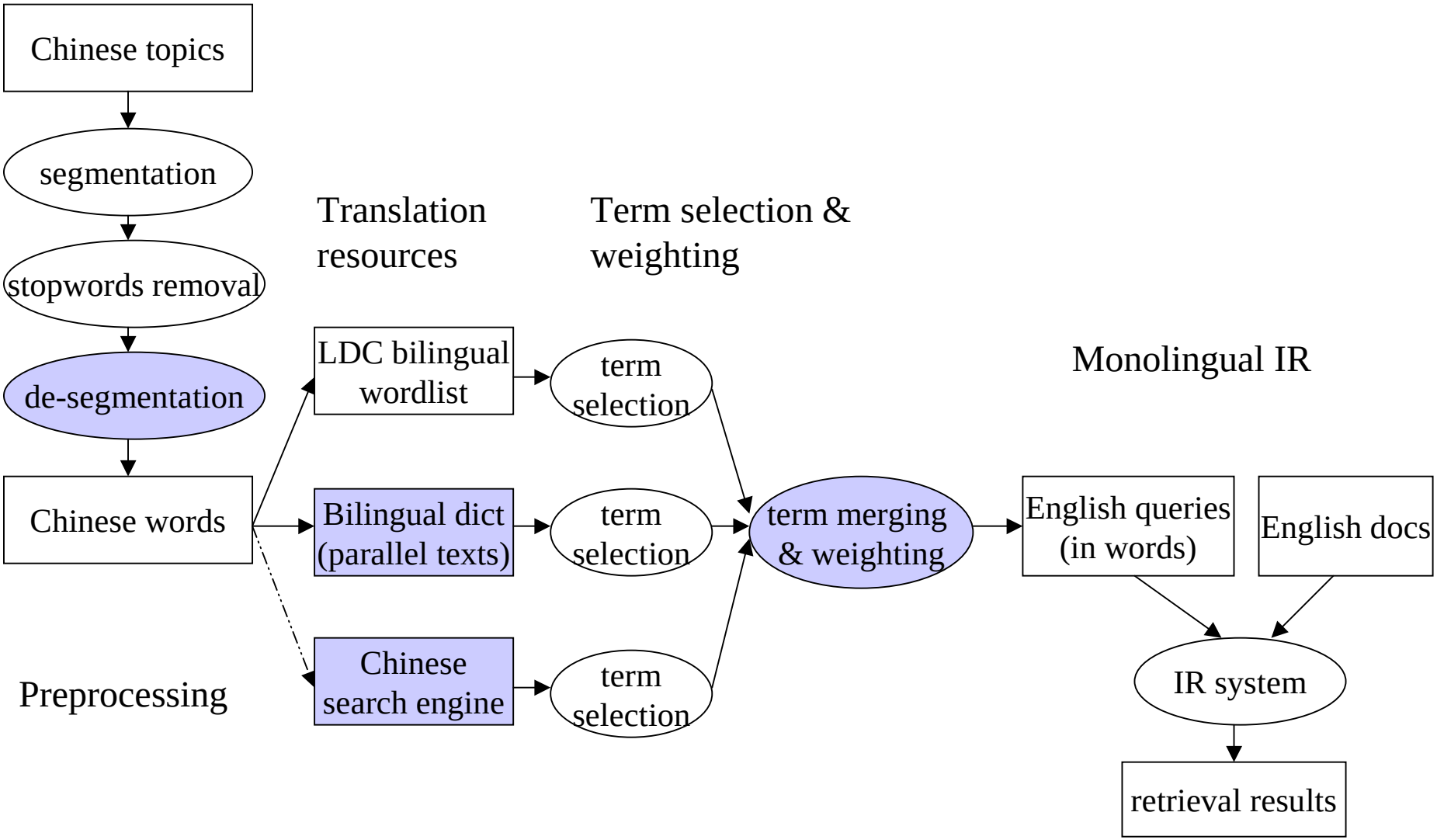
Only component words of compounds are kept in the queries.

German Monolingual Retrieval Performance



Features: +stemming, +decompounding, -expansion

Overview of Chinese to English Retrieval



Chinese Topics Preprocessing: De-segmentation

Topic: C043 El Nino and the Weather

——————— after segmentation ———————

Title: 天氣 與 聖 嬰 現象
Desc : 查詢 有關 聖 嬰 現象 的 說明 ，以及 對 全球 天氣 （ 包括 氣溫 、 氣壓 、 降雨 等 ）的 影響 。

——————— after desegmentation ———————

Title: 天氣 與 聖嬰(El Nino) 現象
Desc : 查詢 有關 聖嬰(El Nino) 現象 的 說明 ，以及 對 全球 天氣 （ 包括 氣溫 、 氣壓 、 降雨 等 ）的 影響 。

Topic: C044 Indurain Wins Tour

——————— after segmentation ———————

Title: 英 杜 蘭 贏得 冠軍
Desc : 有關 米格 爾 · 英 杜 蘭 贏得 第四 屆 環 法 自由 車賽 的 反應 。

——————— after desegmentation ———————

Title: 英杜蘭(Indurain) 贏得 冠軍
Desc : 有關 米格 爾 · 英杜蘭(Indurain) 贏得 第四 屆環法(Error) 自由 車賽 的 反應 。

Translation Resources: Creation of Bilingual Dictionary From Parallel Texts

- Parallel texts: Hong Kong news (4/98-4/2001) and FBIS Chinese collection.
- Document alignment: IR + LDC wordlist.
- Paragraph & sentence alignment: adapted from Gale and Church's length-based model.
- Association measure: Dunning's maximum likelihood ratio statistic.

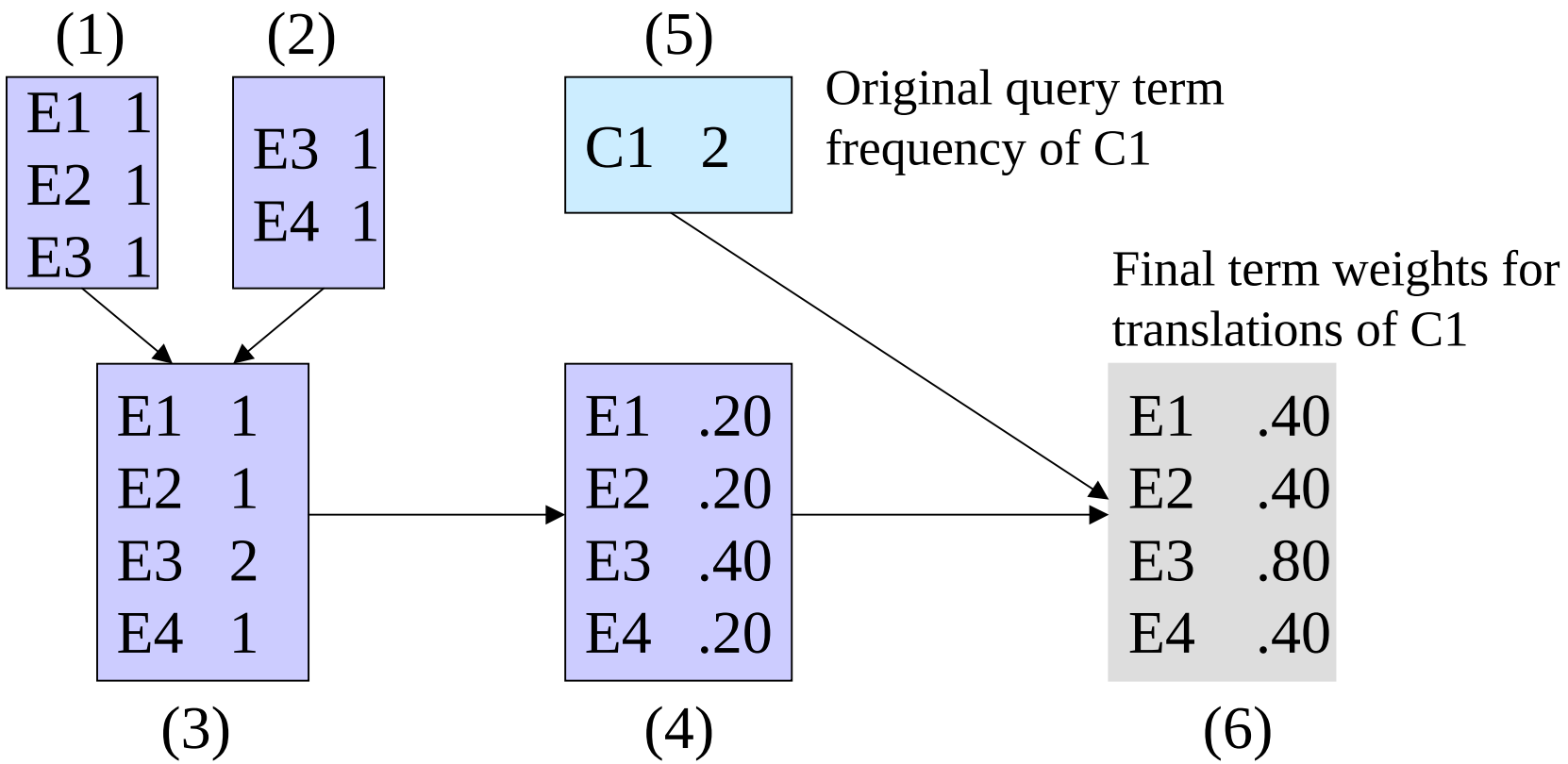
Term Translation Using Search Engine

Search Term	Yahoo.China Search Results
狂牛症	... 因為狂牛症（bovine spongiform encephalopathy）（簡稱 BSE）於1980年末出現了 ..
	... 近一兩個月德國狂牛症(BSE)有蔓延之勢 ...
	... 年紀最大的變種狂牛病（mad cow disease）患者後，...
	... 何謂「狂牛症」(mad cow disease, Bovin Spongiform Encephalopathy,BSE)？...
	... 狂牛症(Mad-cow Syndrom). ...
英杜蘭	... 力圖成為繼95年英杜蘭(Miguel Indurain)之後， ...
尤利西斯	... 太陽探測船—尤利西斯號（Ulysses）...
	... 斯走的途徑大致追隨著尤利西斯（Ulysses）的漂流路線，...
	... 決派尤利西斯(Ulysses)找尋阿基里斯。 ...
	... 如 James Joyce 的尤利西斯 (Ulysses) ，典故就是史詩奧德賽 (Odyssey)；...
綠松石	... 綠松石(Turquoise): 它是十二月的誕生石。
	... 天然鹼(Trona)*. 綠松石(Turquoise).釩鈣鈾礦(Tyuyamunite). 釩鉛礦(Vanadinite). ...
	... 綠松石的英文名稱為Turquoise，源于法語Pierreturqoise，意思是"土耳其石"...
藍濃	... 搖滾樂後來許多發展則可追溯到藍濃（ John Lennon ）的"TomorrowNever Knows"...
藍儂	...（Beatles）的創作靈魂 - 約翰·藍儂（JohnLennon）和其愛侶小野洋子的傑作。
	... 已故前披頭士合唱團主唱約翰藍儂(John Lennon)儘管已經逝世廿年 ...
	... 而披頭四成員之一的約翰.藍儂（John Lennon）1980年在這棟公寓裡遇刺， ...

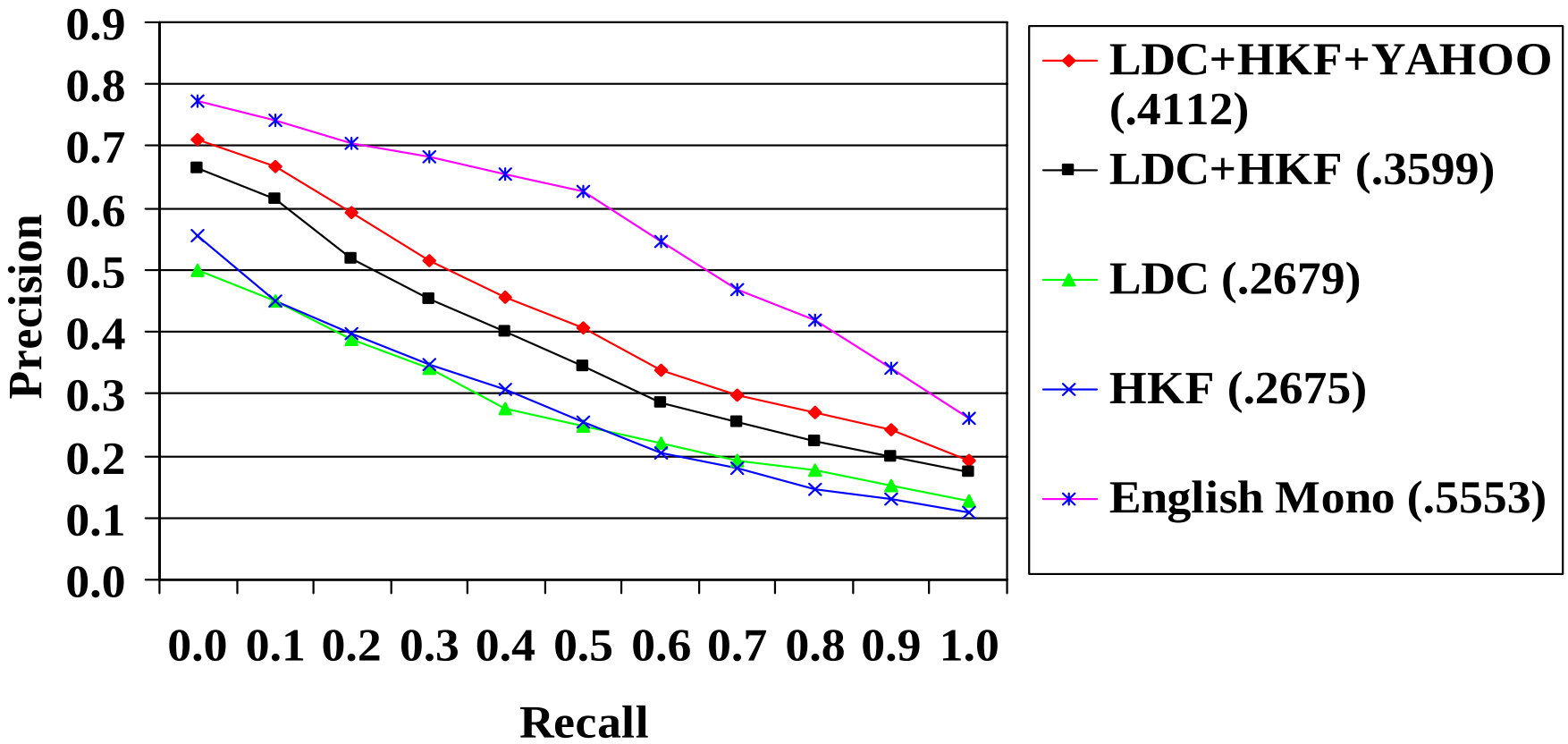
Term Selection, Merging, and Weighting

(1) Top-3 English translations of Chinese word ‘C1’ from LDC wordlist. Translations are ranked by occurrence frequency in the LA Times collection.

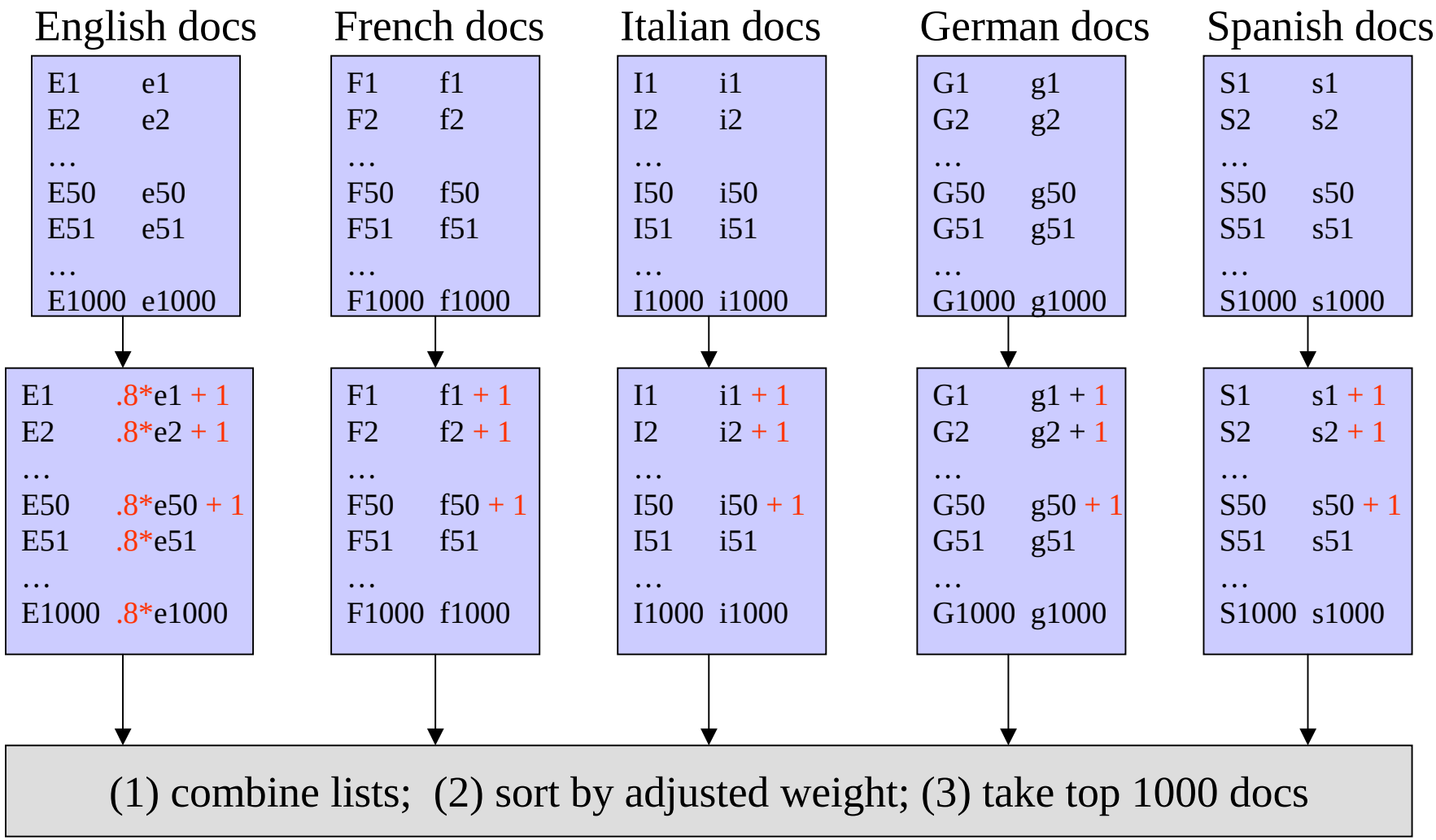
(2) Top-2 English translations of Chinese word ‘C1’ from parallel texts. Translations are ranked by association weight.



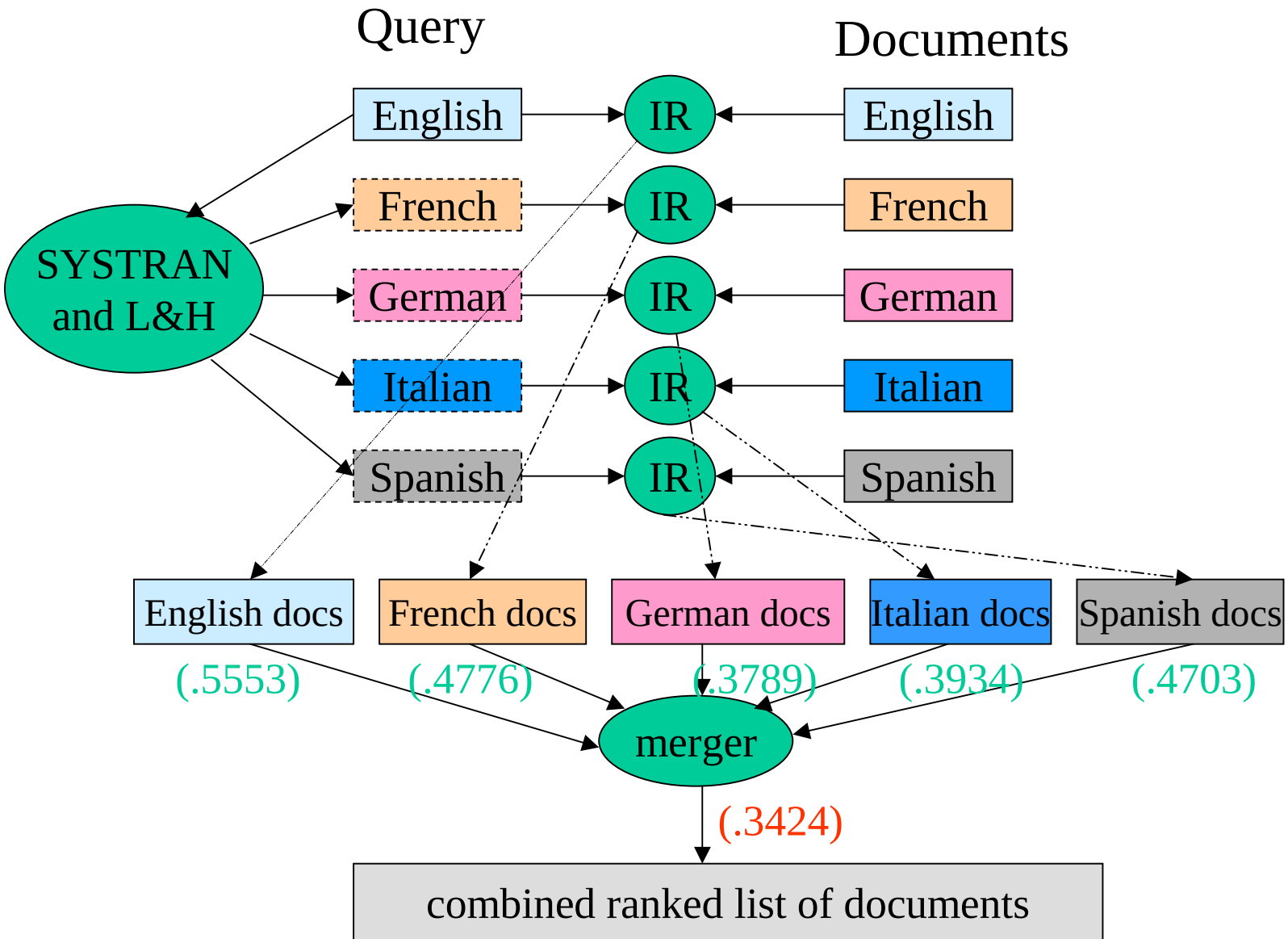
Translation Resources Versus Chinese-to-English IR Performance



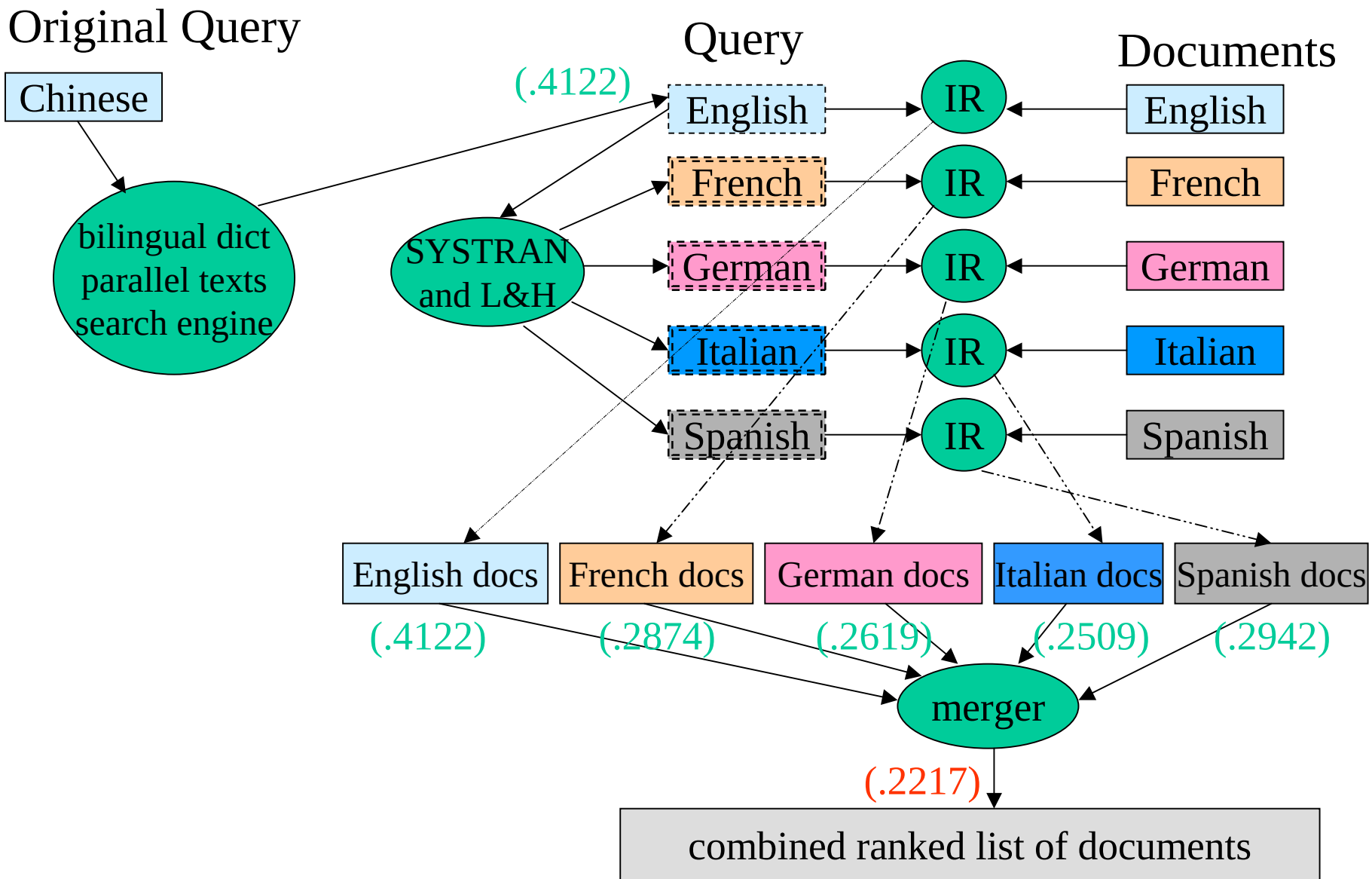
Multilingual Information Retrieval: Merging Strategy



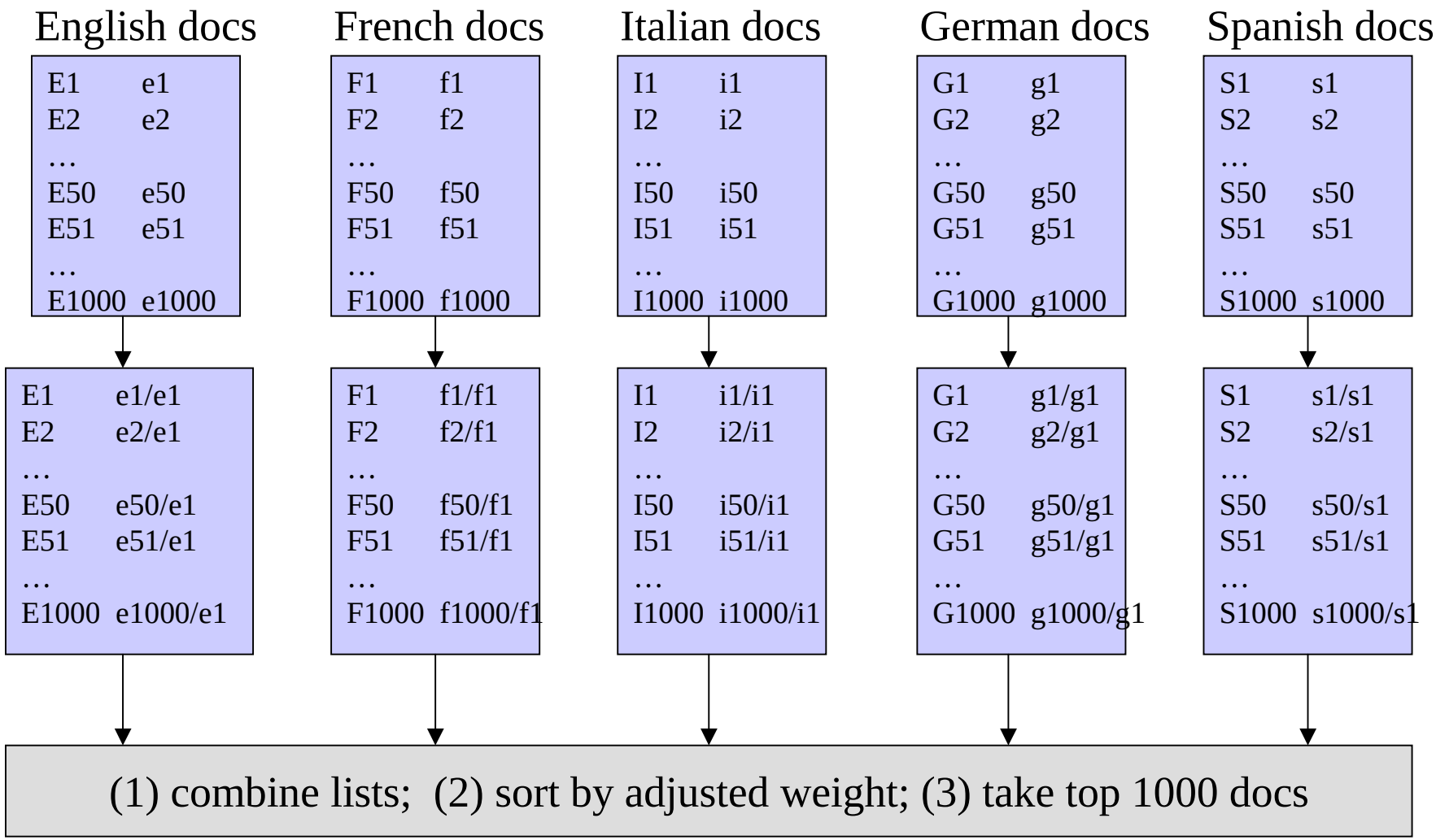
Performance of Multilingual Information Retrieval Using English Long Queries



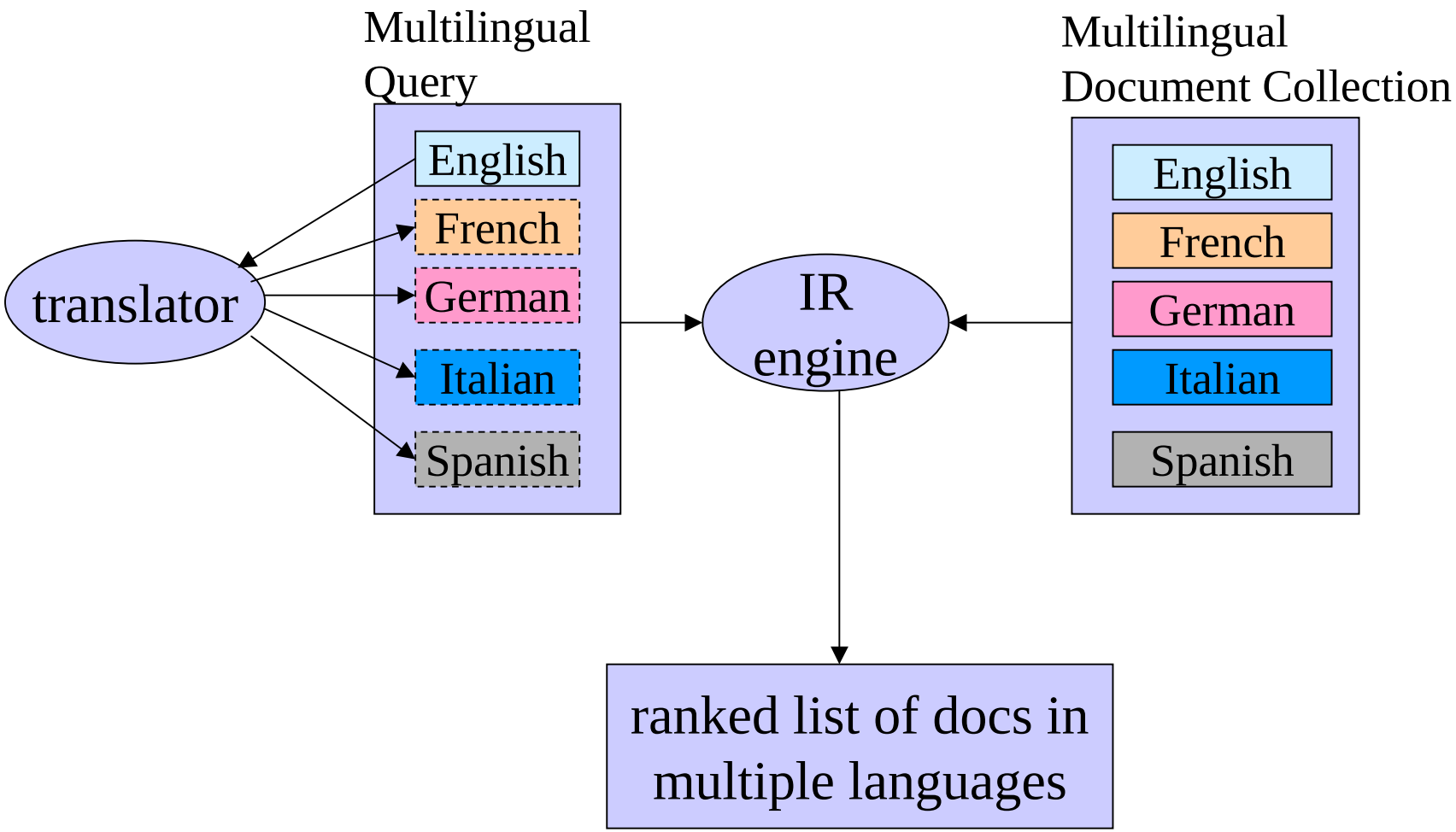
Performance of Multilingual Information Retrieval Using Chinese Long Queries



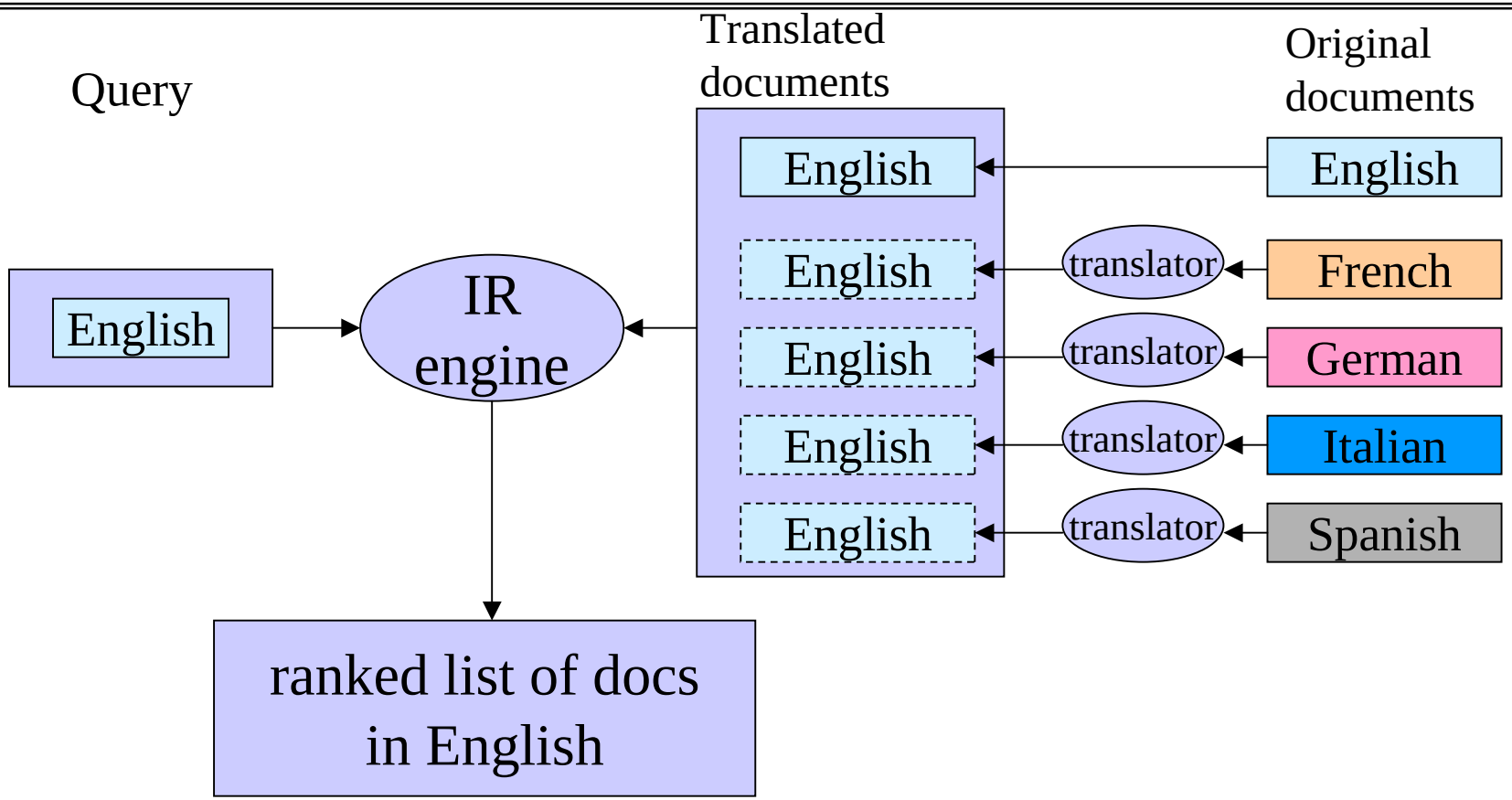
Multilingual Information Retrieval: Alternative Merging Strategy



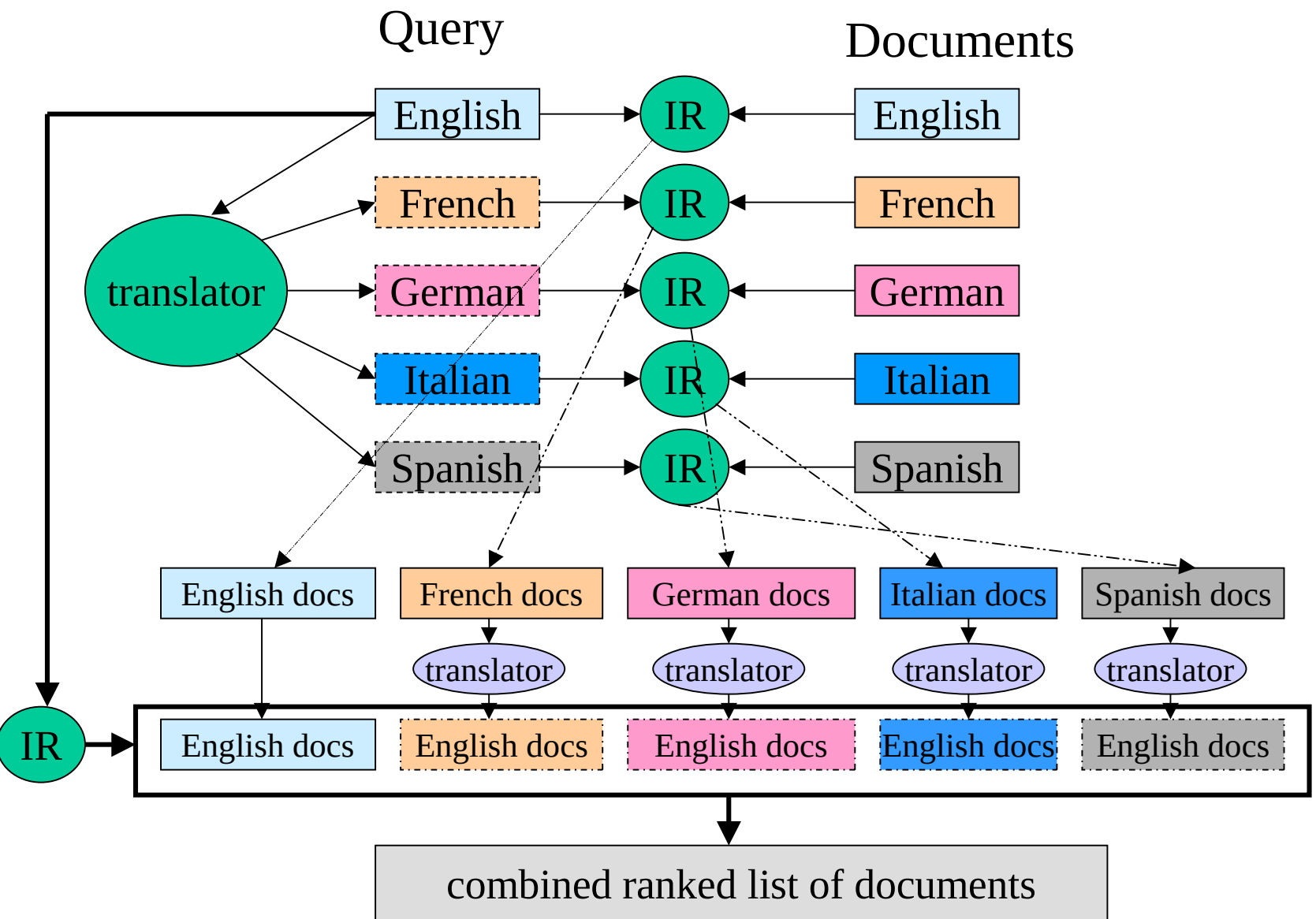
Multilingual Information Retrieval: Alternative Method 1



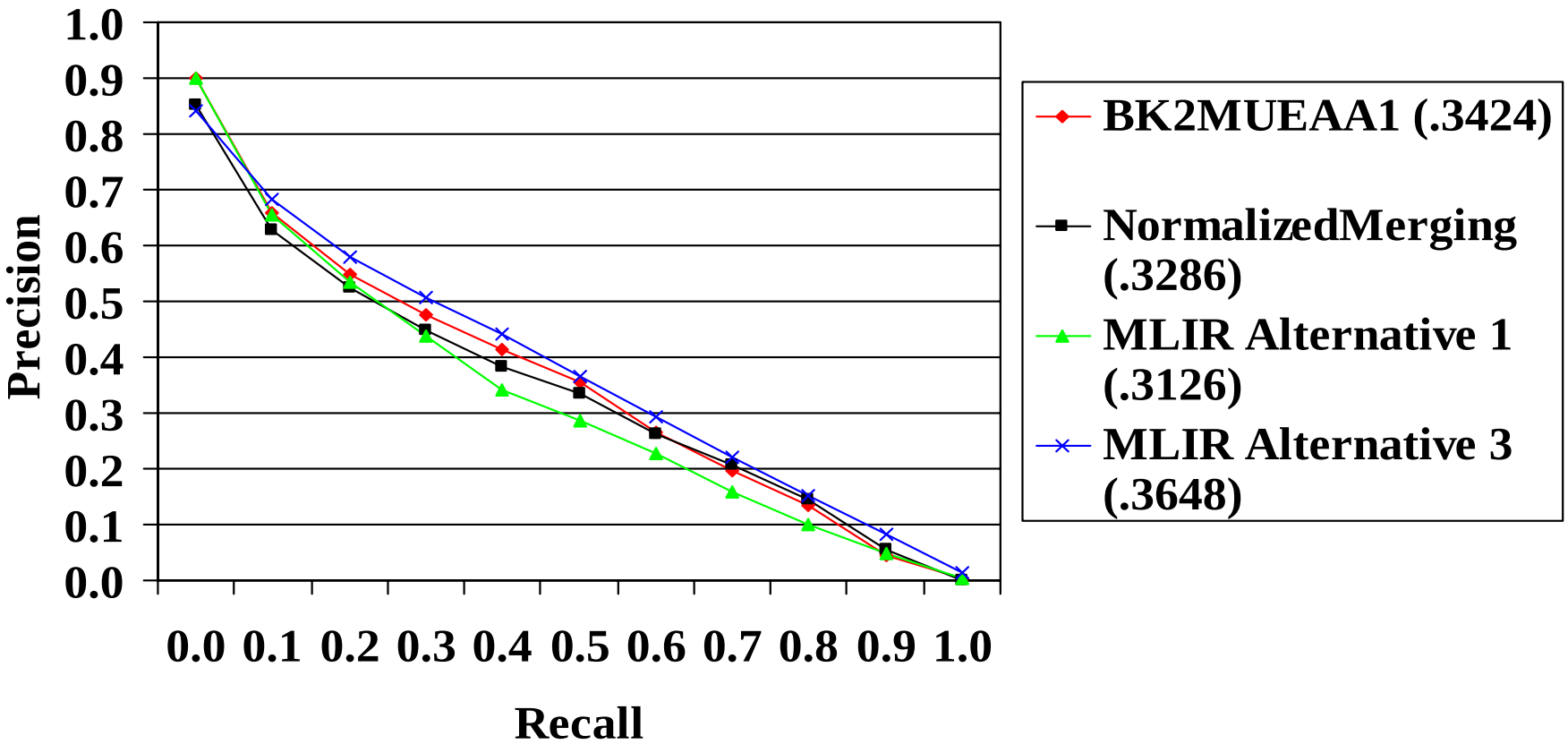
Multilingual Information Retrieval: Alternative Method 2



Multilingual Information Retrieval: Alternative Method 3



Performance of Different MLIR Methods



Conclusions

- German decompounding can significantly improve retrieval performance. Keeping only component words in the query works better than keeping both compounds and component words.
- Chinese search engine is a valuable resource for translating Chinese proper nouns into English.
- Merging documents by adjusted probability of relevance works reasonably well.