

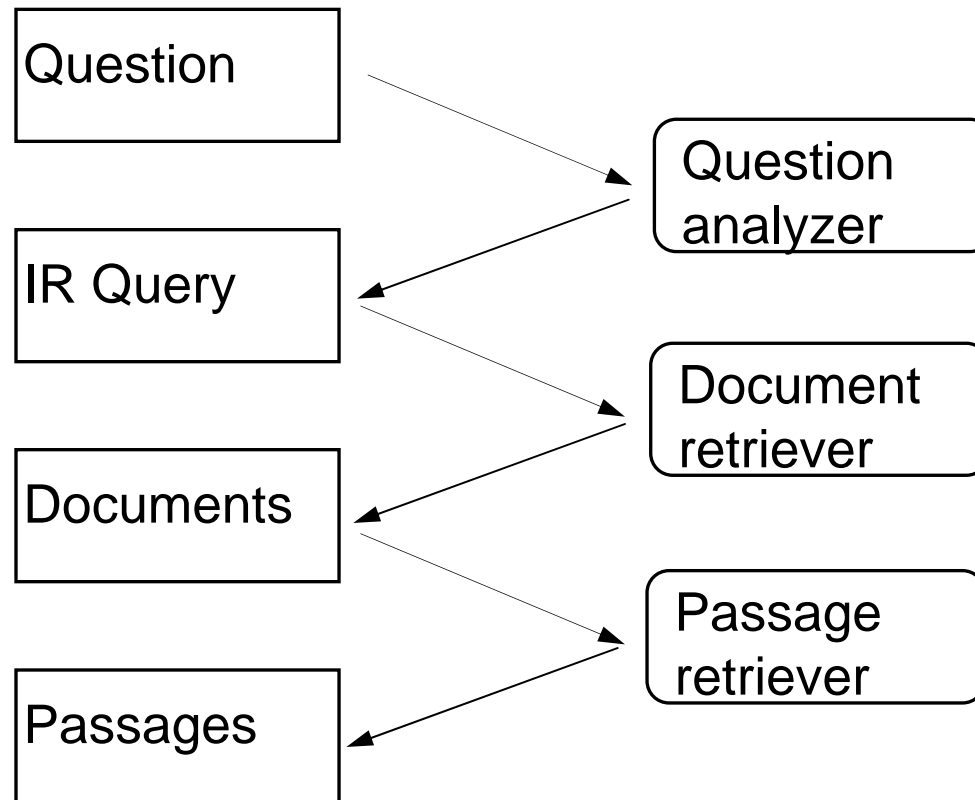
# UIC at TREC 2007: Genomics Track

Wei Zhou, Clement Yu

University of Illinois at Chicago

Nov. 8, 2007

# Overview of the system



# Interpretation of a query

- Q: “What [ANTIBODIES] have been used to detect protein TLR4?”

TARGET: an instance of the entity type

QUALIFICATION

A **TARGET** refers to any instance of the entity type

The **QUALIFICATION** refers to the condition that the target needs to satisfy in order to be qualified as an answer to the query

# Two types of queries

Type I: candidate targets can be found in existing resources.

SIGNS OR SYMPTOMS
Headache
Coughing
.....

~ UMLS

GENES
CD44
PSD-95
.....

~ Entrez Gene

Type II: no resources available to find candidate targets.

“ANTIBODIES”, “PATHWAYS”, and etc.

Type I		
ENTITY TYPE	Resource	Mapping
BIOLOGICAL SUBSTANCES	UMLS	Biologically Active Substance
CELL OR TISSUE TYPES	UMLS	Cell and Tissue
DISEASES	UMLS	Disease or Syndrome
DRUGS	UMLS	Pharmacologic Substance
GENES	Entrez Gene	Gene symbols
MOLECULAR FUNCTIONS	UMLS	Molecular Function
PROTEINS	UMLS	Amino Acid, Peptide, or Protein
SIGNS OR SYMPTOMS	UMLS	Sign or Symptom
TUMOR TYPES	UMLS	Neoplastic Process

Type II
ENTITY TYPE
ANTIBODIES
MUTATIONS
PATHWAYS
STRAINS
TOXICITIES

# IR models for Type I queries

Observation: different passages may have different qualified targets.

Strategy I: The more important qualified targets a passage has, the higher it is ranked (the importance is measured by idf).

Strategy II: Passages having at least one qualified target are ranked equally.

# IR models for Type I queries

- Strategy I:

$$sim(q, d) = (sim(q, d), sim(q, d))$$

$$sim(q, d) = \alpha \times \underset{\text{concept}}{MAX} \{ \underset{\text{word}}{wt(g_i)} \mid g_i \in d \} + (1 - \alpha) \times \sum_{c \in d} wt(c)$$

- Strategy II:

$$sim(q, d) = (sim(q, d), sim(q, d), sim(q, d))$$

target                  qualification                  word

$$\underset{\text{target}}{sim(q, d)} = 1, \text{ if } d \text{ has at least one target}$$
$$= 0, \text{ otherwise}$$

# A conceptual IR model

$$\text{sim}(q, d) = [\text{sim}(q, d), \text{sim}(q, d)]$$

concept word

$\text{sim}(q, d_2) > \text{sim}(q, d_1)$  if either 1 or 2: (Liu 2004)

1.  $\text{sim}(q, d_2) > \text{sim}(q, d_1)$

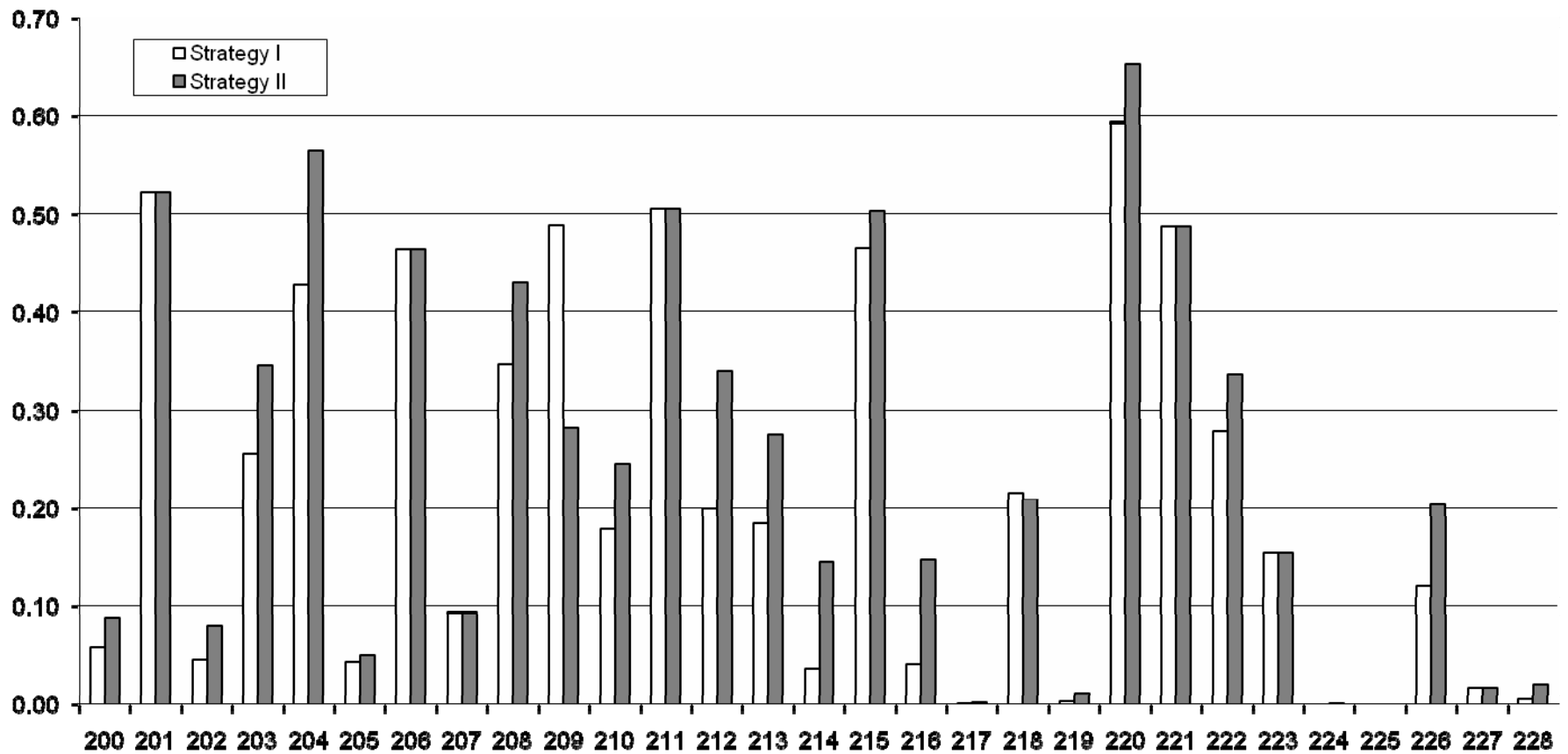
concept concept

2.  $\text{sim}(q, d_2) = \text{sim}(q, d_1)$  AND  $\text{sim}(q, d_2) > \text{sim}(q, d_1)$

concept concept word word

# Result: Strategy I vs. Strategy II

	Doc MAP	Asp MAP	Psg1 MAP	Psg2 MAP
Strategy I	0.209	0.145	0.083	0.044
Strategy II	0.239	0.181	0.098	0.051



# IR models for Type II queries

- Query expansion with representative words of the entity type.

Representative words: words that co-occur with the entity type much more often than being independent in the same sentences of PubMed abstracts. The top 15 representative words were added into the query. For example:

ANTIBODIES: “antigen”, “humanized”, “neutralizing”, “vaccine”, “cell”, “immune”, “protein”, “assay”, “immunogenicity”, “vaccination”, and etc.

Similarity measure:  $sim(q, d) = (sim(q, d), sim(q, d)')$   
concept word

## Result: *candidate targets vs. representative words*

	Doc MAP	
	Candidate targets	Representative words
Median	0.1659	0.2704
UICGene2	0.2147	0.3410
Diff.	29.42%	26.11%

The representative words of entity type is useful: they have captured the context in which a target of that entity type appears.

# Passage extraction

## **A two-step rational:**

Step 1: Given various windows of different sizes, choose the ones which have the **maximum** number of query concepts and the **smallest** number of sentences.

Step 2: Merge two candidate windows if they are exactly adjacent to each other.

# Experience of using UMLS

1. Some candidate targets retrieved from UMLS are very common terms and they are not “real” instance of the entity type.

Candidate Target	UMLS semantic type	ENTITY TYPE
Water	Pharmacologic Substance	DRUG
Tumor	Neoplastic Process	TUMOR TYPES
Brain	Disease or Syndrome	DISEASES

# Experience of using UMLS

2. Some candidate targets are variants of the entities in the documents.

A sample relevant passage:

“In patients with cad, several small clinical trials suggest that cognitive-behavioural therapy successfully reduced **anxiety** and depression, and thus facilitated the modification of cardiac risk factors.”

Sign or Symptom (UMLS)

grade 1 anxiety

grade 2 anxiety

grade 3 anxiety

grade 4 anxiety

grade 5 anxiety

# Summary

- 1. A better performance has been achieved to rank passages having at least one qualified target equally than to rank them according to the importance of targets they have.
- 2. A filtering step might be necessary to retrieve candidate targets from UMLS.

- **THANK YOU**