Text Processing on the Web

Week 6

Query Expansion and Passage Retrieval

The material for these slides are borrowed heavily from the precursor of this course taught by Tat-Seng Chua as well as slides from the accompanying recommended texts Baldi et al. and Manning et al.

Recap: PageRank and HITS

- Using only the social network of directed hyperlinks to do ranking

- Holistic: Pagerank
  - Prestige via Random Walk
- Dualistic: Hubs and Authorities

What's their intrinsic relationship?

Three-week Outline

Today

- External Resources
  - Thesaurii
  - Wikipedia
  - Domain specific Sites
- Query Expansion
  - Query logs to suggest
- Ranking
  - Density Based
  - Dependency Based

Next Time

- What is Question Answering?
  - TREC
  - Def, List, Factoid, OpEd, Event
  - Closed vs. Open Domain
- Question Analysis
  - Question Typologies
- Refining from Passages
  - Answer Justification

Outline

Heading towards exact answer retrieval (to be examined in detail after recess week)

Today (waypoint): passage retrieval

Also: Using external resources

During break: Mid-term feedback

What is passage retrieval?

Retrieving passages instead of full documents

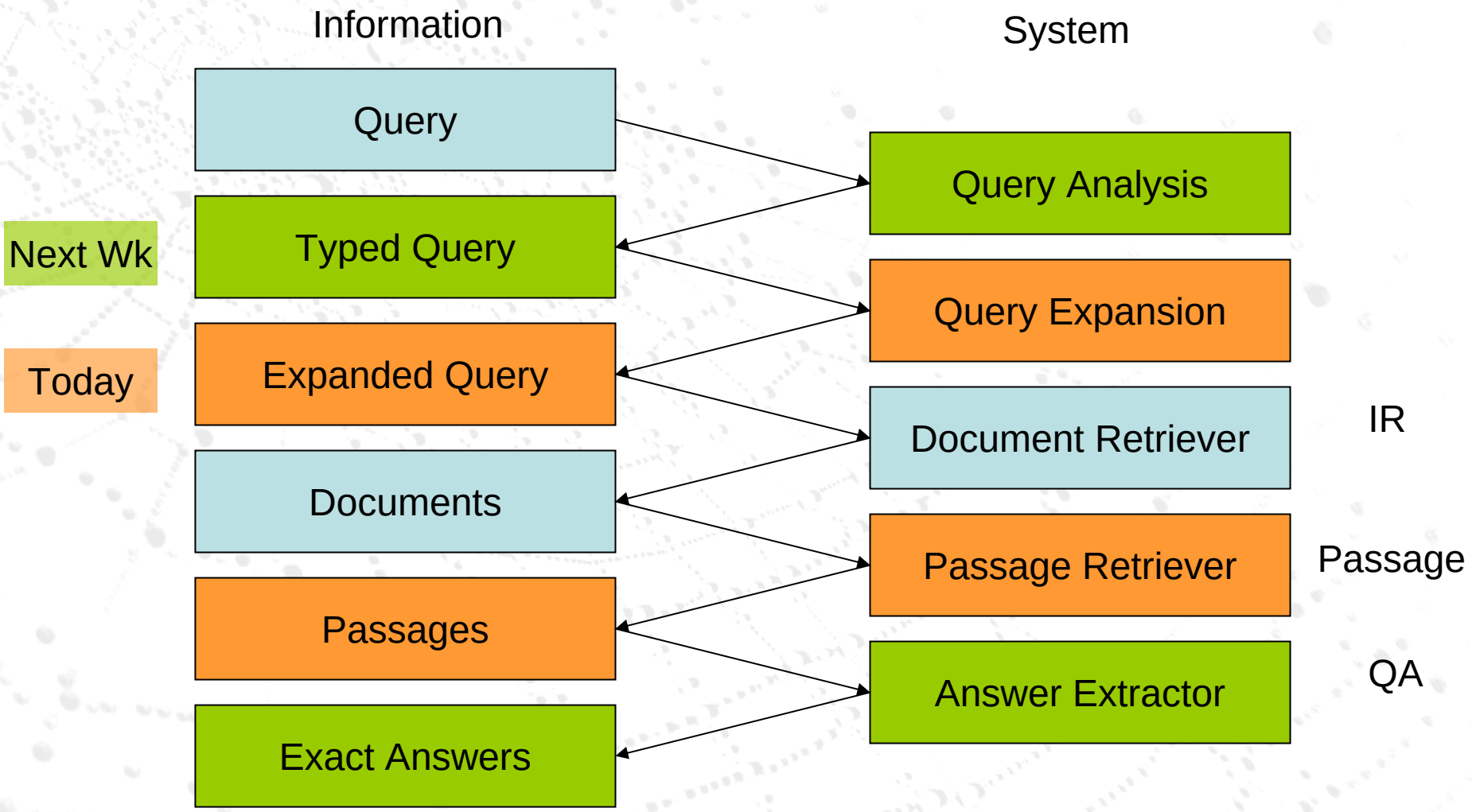
- What are passages?
  - Could be sections, paragraphs or sentences where sections delimited by format (analysis)
  - Size limited definition: up to n words or bytes (snippets)
- Why?
  - Zoom in on answer
  - Information retrieval vs. document retrieval

The importance of context

Context grows in importance as we approach exact question answering. Why?

- Documents are usually independent, stand alone, fully interpretable units
  - On the web: what are exceptions to this?
- Passages are usually not, need context to properly interpret
  - Again, why are exceptions to this?
- Passages are also harder to rank due to their smaller size

Passage Retrieval Architecture



Architecture notes

- Query Expansion
  - To overcome mismatch between query and target text
  - May use external resources, this can lead to performance gains (over 10% in F_1)
- Passage Retrieval
  - Ranking (or re-ranking of candidate passages)

Uses a combination of heuristic and machine learning approaches

Still much more to do here, no coherent framework for this work yet

Query Expansion

We've already seen Rocchio (for relevance feedback)

How do you do expansion without feedback?

One answer: Pseudo Relevance Feedback (PRF) -does help (as most queries are about the main sense of words)

These are internal to the document collection. What about external resources?

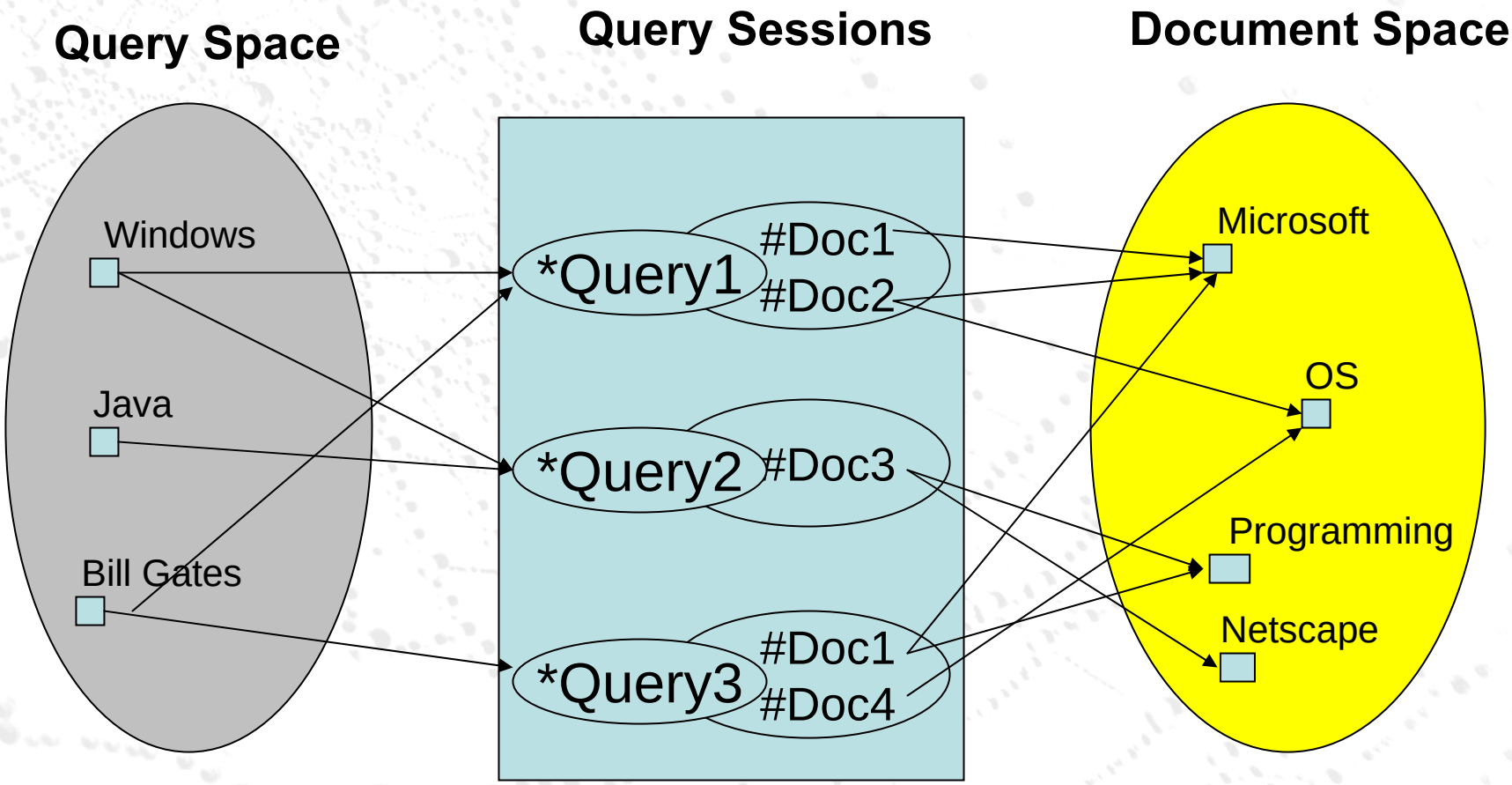
- Users: Query Logs
- Knowledge bases: Concept taxonomies, Lexical databases
- Web: Content terms from websites

Probabilistic Query Expansion Using Query Logs

- MS Encarta Encyclopedia study with 41K docs, 4M queries
- Big gap between document and query space
  - Cosine sim: around .1-.4 ;Average angle: 73 deg
  - Lots of doc words never used in query, need to filter

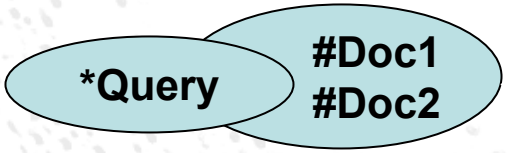
- Answer: Use query expansion from query logs

Query sessions as a bridge



Term-term correlation

Query Term



Index Term

Relevance of query term to a doc

$$P(D_k \mid w_i^{(q)}) = \frac{f_{ik}^{(q)}(w_i^{(q)}, D_k)}{f^{(q)}(w_i^{(q)})}$$

Relevance of doc to an index (doc) term

$$P(w_j^{(d)} \mid D_k) = \frac{W_{jk}^{(d)}}{\max_{\forall t \in D_k}(W_{tk}^{(d)})}$$

Stuff them together

$$P(w_j^{(d)} \mid w_i^{(q)}) = \sum_{\forall D_k \in S} (P(w_j^{(d)} \mid D_k) \times \frac{f_{ik}^{(q)}(w_i^{(q)}, D_k)}{f^{(q)}(w_i^{(q)})})$$

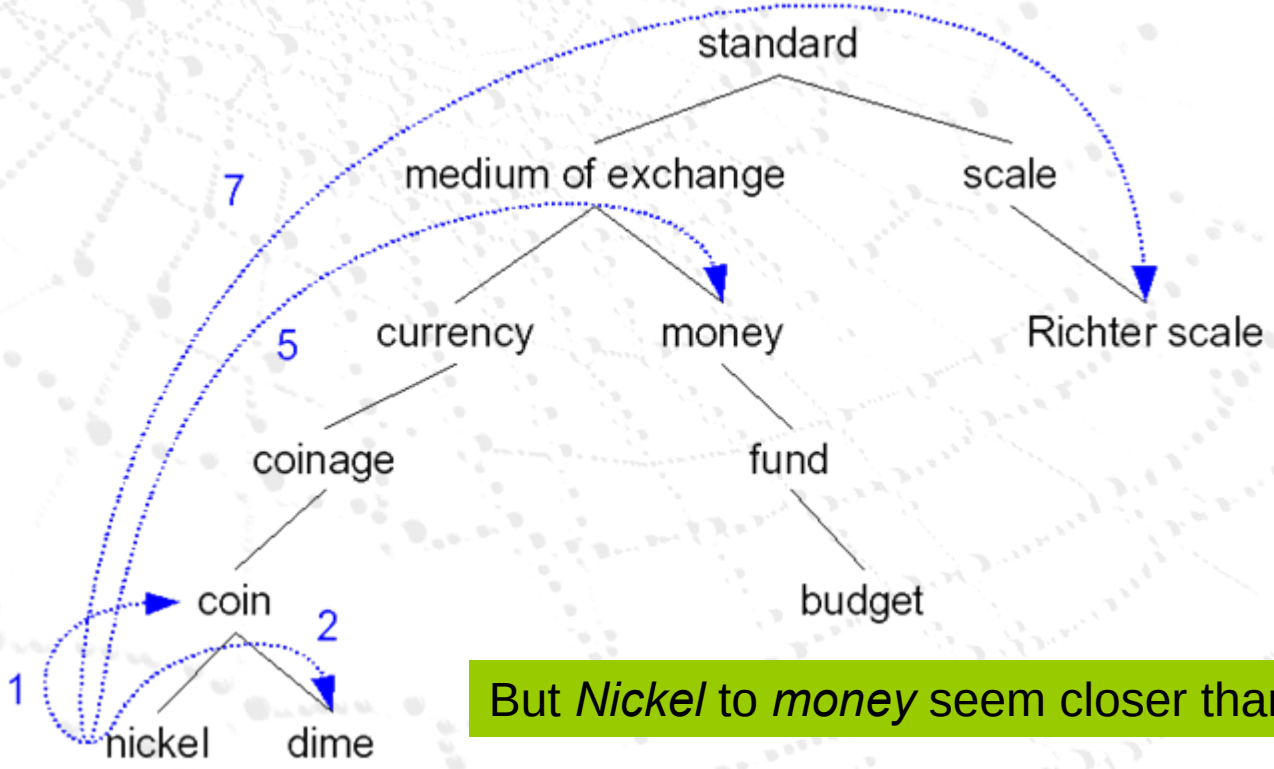
Lexical Database

- Common in closed domains where a limited vocabulary is used (controlled vocabulary)

- Has broader and narrower terms
- Related or synonymous terms
- Example: WordNet (www.cogsci.princeton.edu/~wn/)
  - Organizes hierarchy around a *synset*
  - i.e., synonym set -- a group of tokens that can express the same core concept
  - e.g., (synset #07754049) *spring, fountain, outflow, outpouring, natural spring*

Path based similarity

- Two words are similar if nearby in thesaurus hierarchy (i.e. short path between them)



But *Nickel* to *money* seem closer than *nickel* to *standard!*

Information content similarity metrics

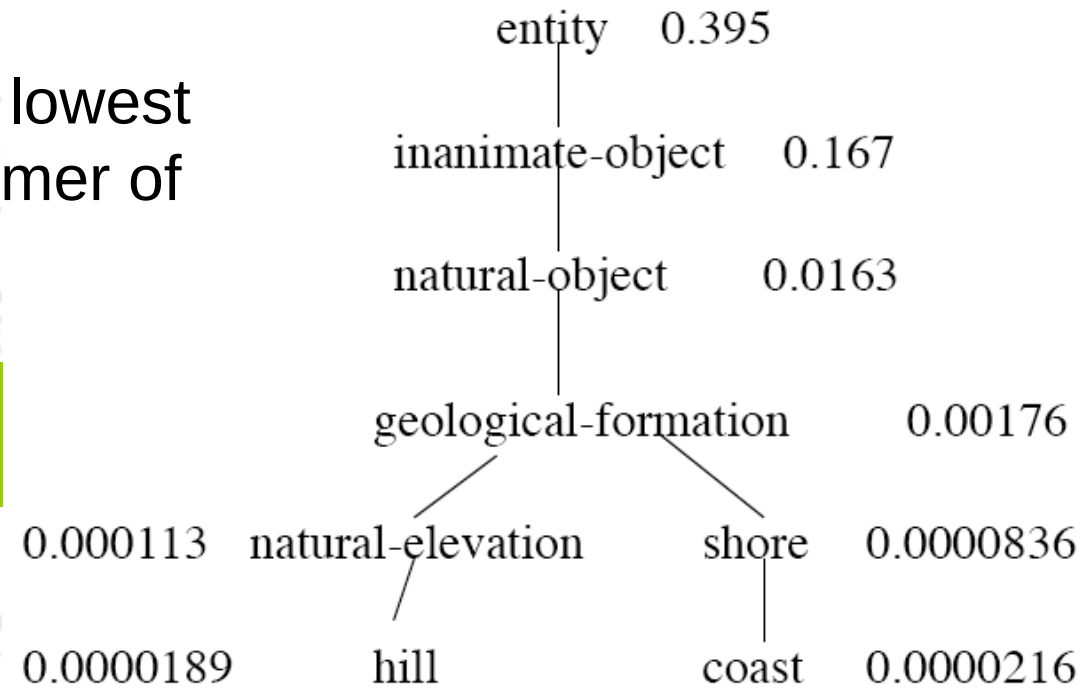
- Let's define P(C) as:
  - The probability that a randomly selected word in a corpus is an instance of concept *c*
  - Formally: there is a distinct random variable, ranging over words, associated with each concept in the hierarchy
  - P(root)=1
  - The lower a node in the hierarchy, the lower its probability

Information content similarity

How to use these probabilities?

One way: find the lowest common subsumer of terms a & b

How should we obtain these probabilities?



WordNet hierarchy augmented with probabilities P(C)

Information content: definitions

- Information content:

  $IC(c) = -\log P(c)$
- Lowest common subsumer

  $LCS(c_1,c_2)$ = the lowest common subsumer
  - I.e. the lowest node in the hierarchy
  - That subsumes (is a hypernym of) both c_1 and c_2
- We are now ready to see how to use information content IC as a similarity metric

Using Web Resources

- Extract correlated terms from general or specific web resources
  - General: Wikipedia, SE snippets, Dictionary
  - Specific: IMDB, Rotten Tomatoes, MovieLens, NetFlix

- Use the Web as a generalized resource:
  - Retrieve N docs using Q_0
  - Extract nearby non-trivial words in window as w_i
  - Rank w_is by correlation with query using **mutual information**
  - Incorporate w_is above threshold to be part of Q_1

Passage Retrieval

Word overlap and density methods

Dependency relations

Passage Retrieval

- As stated, an intermediary between documents and answers
- Which is more important for question answering: precision or recall?

A simple method:

- View document as set of passages
- Use some basic IR techniques to rank

Comparing Passage Retrieval

- Tellex et al. (2003) compared a fair number of algorithms

We'll examine:

- Word overlap
- Density based
- External sources

Word Overlap (MITRE)

- Count # of terms in common with question in the passage

- What is this equivalent to in terms of document retrieval?
- Works surprisingly well but wouldn't work well for document retrieval. Why?

Density Based

Look at the density of overlapping terms in the passage

- Favor passages with many terms with high idf value
  - Passages must start and end with a query term (**Multitext**)
  - Passages are *n*-sentences long (n=3, **SiteQ**)
  - Consider also adjacency (n-grams, **IBM**)

External Resources

- Thesaural Relations (**IBM**)
- Named Entity Tagging (e.g., separate match score for names of people, organizations and places; **ISI**)
- Web based expansion, head nouns, quotation words (**NUS**)

Usually:

- Linearly combined together with overlap or density based metrics (IBM, ISI)
- Or cascaded in iterative successions (NUS)

Which works best?

- Tellex et al. tested variants of these systems along with a committee voting algorithm

	Mean Reciprocal Rank (MRR)	
System	**Exact**	**Lenient**
Multitext	.354	.428
IBM	.326	.426
ISI	.329	.413
SiteQ	.323	.421
Alicante (Cosine based)	.296	.380
BM25 (Prob IR)	.312	.410
MITRE (Word Overlap)	.271	.372

So what does this mean?

- Density-based methods perform better than simple overlap
  - Non-linear boost for terms that occur close to each other
- Passage retrieval algorithms achieve higher MRR when using simple Boolean document retrieval
  - Let passage retrieval do the refinement
  - Aim instead for higher recall

What else?

Light et al. observed that up to **12**% of answers in certain dataset have no overlap at all with the question

- Shows value in term expansion

- But term expansion noisy; need to filter out incorrect expansions
  - Especially in the context of the web
  - Especially since passage retrieval should emphasize precision

Density Based Passage Retrieval Method

- Density based methods can err when …

<Question> What percent of the nation's cheese does Wisconsin produce?

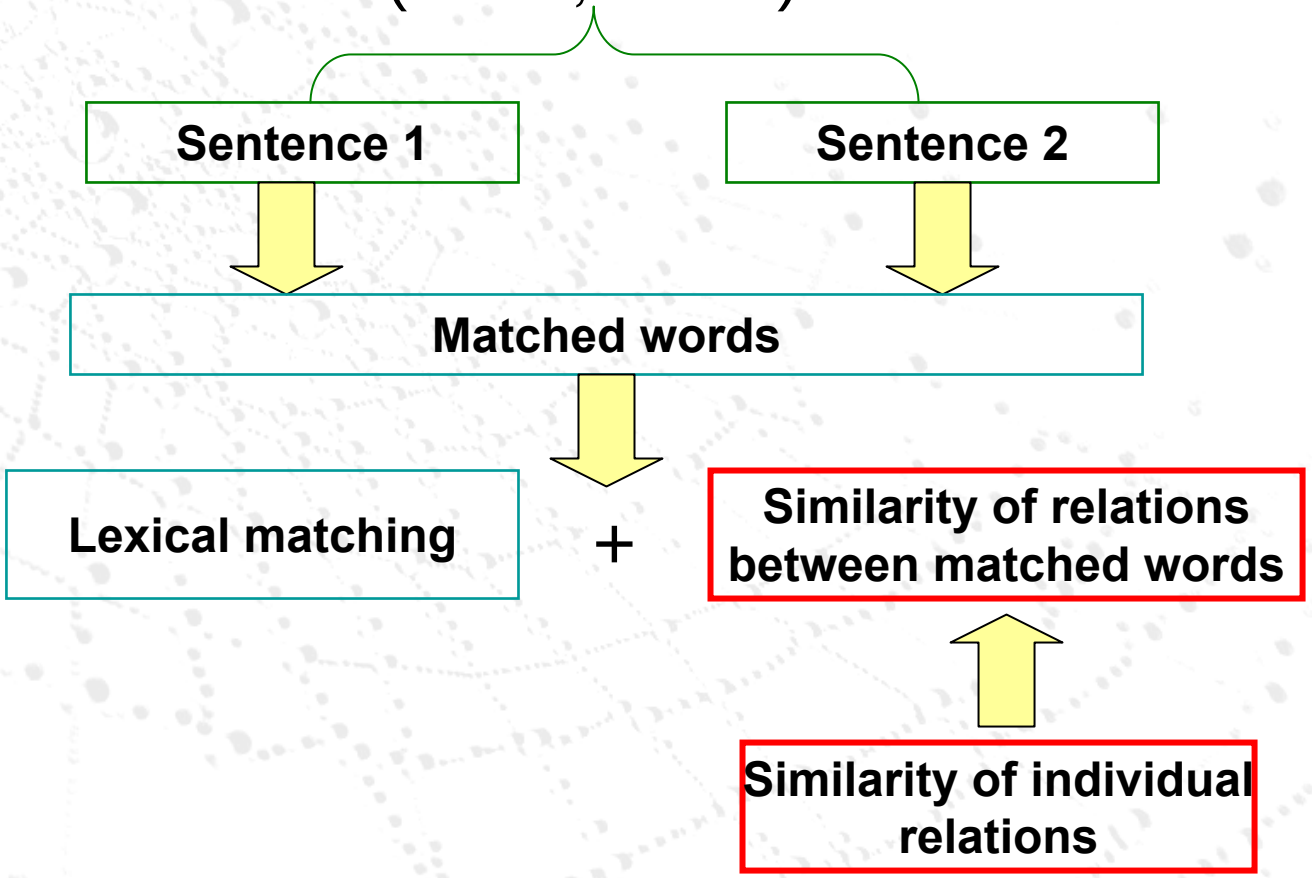
Incorrect: … the number of co
about *cheese* has risen by 14
has dropped 16 *percent*.
Incorrect: The wry "It's the *Che*
its *cheese* _ and indulge in an
of *cheese* in California grew three time
whole 3.7 *percent* compared to 1.2
Incorrect: Awareness of the Real California *Cheese* logo, which appears on about 95 *percent* of California *cheeses*, has also made strides.

Correct: <S1> In *Wisconsin*, where farmers *produce* roughly 28 *percent* of the *nation*'s *cheese*, the outrage is palpable.

Relationships between matched words differ …

Measuring Sentence Similarity

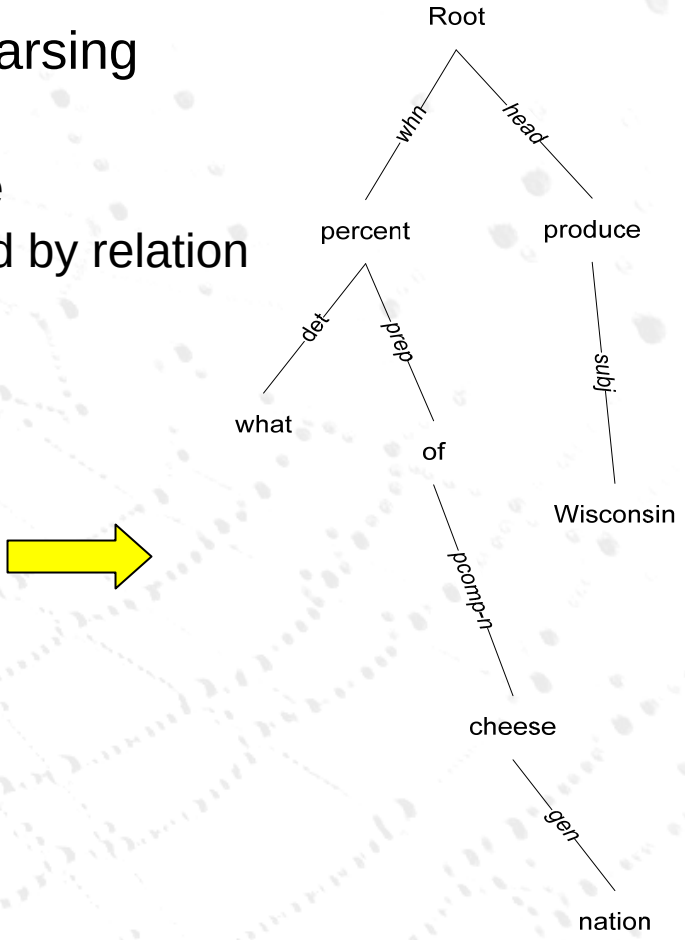
Sim (*Sent1*, *Sent2*) = *?*



What Dependency Parsing is Like

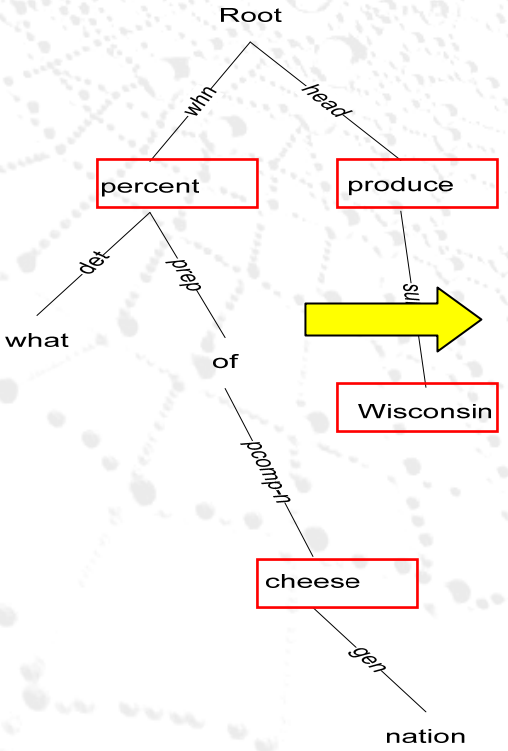
- Minipar (Lin, 1998) for dependency parsing
- Dependency tree
  - Nodes: words/chunks in the sentence
  - Edges (ignoring the direction): labeled by relation types

What percent of the nation's cheese does Wisconsin produce?



Extracting Relation Paths

- Relation path
  - Vector of relations between two nodes in the tree



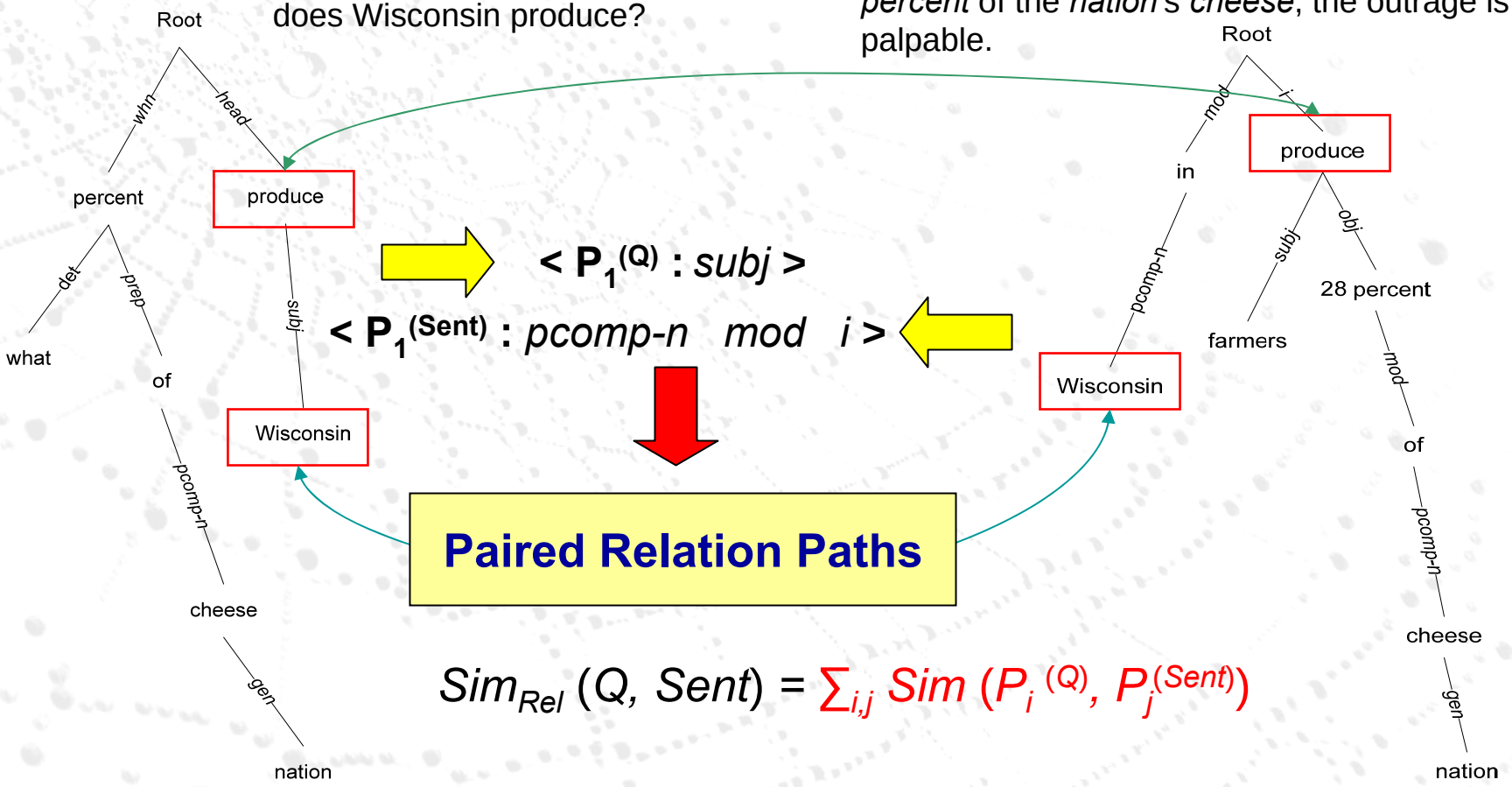
produce **< P_1: *subj* >** Wisconsin

percent **< P_2: *prep pcomp-n* >** cheese

Paired Paths from Question and Answer

What percent of the nation's cheese does Wisconsin produce?

In *Wisconsin*, where farmers *produce* roughly 28 *percent* of the *nation*'s *cheese*, the outrage is palpable.



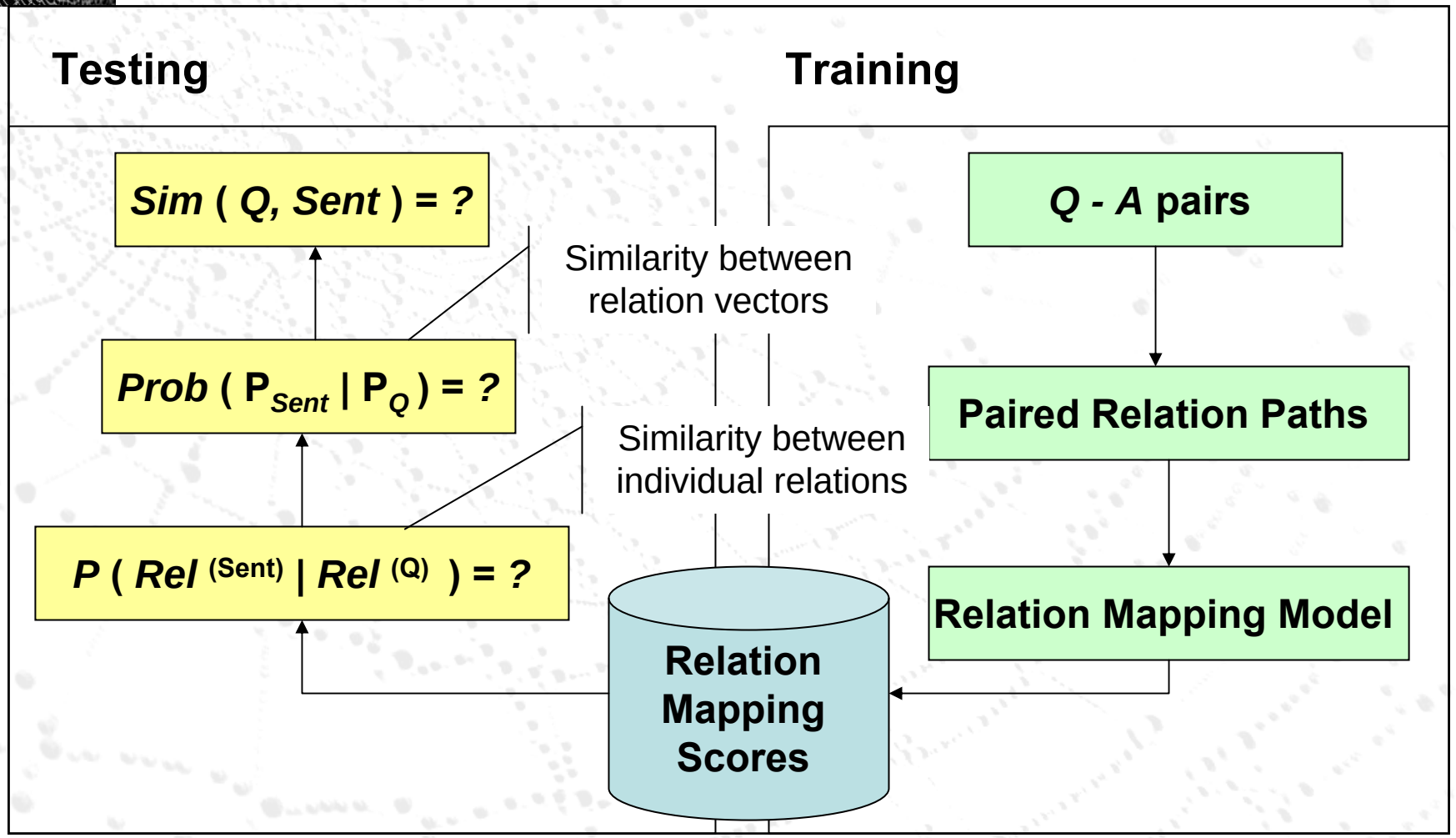
$$Sim_{Rel}(Q, Sent) = \sum_{i,j} Sim(P_i^{(Q)}, P_j^{(Sent)})$$

Measuring Path Match Degree

- But paths don't always match exactly
- Train a similarity method to match non-exact paths

- Path match degree (*similarity)* as a probability
  - *MatchScore* ($\mathbf{P}_Q$, $\mathbf{P}_S$) → *Prob* ($\mathbf{P}_S \mid \mathbf{P}_Q$)
  - *Relations as words*

Training and Testing



Performance

- Helps significantly to use relationships to filter
  - Exact match of paths gives 20% increase in MRR
  - Fuzzy match yields an additional 40%
- Returns best in longer queries. Why?
  - Simple. Longer queries have more noun phrases
  - Noun phrases are used as anchors for paths
  - Thus, more constraints (more data) to help
- Even better performance after query expansion. Why?
  - Again, more input data

Summary

- Tuning the performance of IR systems using
  - Query expansion
  - External resources
- Passage Retrieval
  - Can use simple document methods
  - Are a good platform for trying more substantial processing
  - Emphasizing precision; relegate document retrieval to high recall

References

- **Passage Retrieval Comparison**: S. Tellex, B. Katz, J. Lin, A. Fernandes and G. Marton (2003). Quantitative evaluation of passage retrieval algorithms for question-answering. ACM SIGIR, 41-47
  - **Word Overlap**: M. Light, G. S. Mann, E. Riloff, and E. Breck (2001). Analyses for elucidating current question answering technology. J. Natural Language Engineering, Special Issue on Question Answering, Fall--Winter 2001. http://citeseer.comp.nus.edu.sg/468779.html
  - **Density Based Method**: C Clarke, G Cormack, F Burkowski (1995) Shortest Substring Ranking (MultiText Experiments for TREC-4) Proceedings of TREC-4, 1995