Quantitative Evaluation of Passage Retrieval Algorithms for Question Answering

Stefanie Tellex, Boris Katz, Jimmy Lin, Aaron Fernandes, Gregory Marton
MIT CSAIL (AI + LCS)
Cambridge, Massachusetts, USA

Road Map

- Overview of factoid question answering.
- Our experiments: a quantitative evaluation of passage retrieval algorithms.
- Our findings:
  - Boolean query techniques perform well for question answering.
  - Relative performance of passage retrieval algorithms varies with the document retriever.
  - Density-based scoring drives the best passage retrieval algorithms.

Overview of Factoid Question Answering

- Question answering systems
- Factoid questions
  - “When did Hawaii become a state?” 1959
  - Who was the first American woman killed in the Vietnam War?” Sharon Lane
- Text Retrieval Conference
  - Factoid question answering track in 1999.
  - Formal, rigorous, end-to-end evaluation of question answering systems.

Generic Question Answering System Architecture

- Most TREC QA systems can be decomposed into four components.
  - Question analysis: Decomposes the question for further processing.
  - Document retrieval: Retrieves documents from the corpus.
  - Passage retrieval: Returns paragraph sized chunks from the returned documents.
  - Answer extraction: Returns exact candidate answers.

Question Analysis

When did Hawaii become a state?

- Answer type: date
- Query: Hawaii and become and state
- Proper nouns: Hawaii
- Synonyms
  - Hawaii: HI

Document Retrieval

When did Hawaii become a state?

Tom Selleck Honored by Hawaii Legislature
HONOLULU (AP)

Actor Tom Selleck told lawmakers honoring him
to mark the conclusion
based television serie
it less costly for fi
islands.
Selleck and other
P.I.'' production tea
state Legislature.
In brief remarks be
Selleck said the ``Mag
spent ``$100 million
promoting dollars in
Yet Hawaii's film
competitive because o
said.
One solution is to
for the state's film
making it more attrac
will add more money to
Selleck said.
Charging $25,000 a
in taxes and $1,000 fo
the right signal'' to

American Stock Exchange Plans Trading Facility in
Hawaii
NEW YORK (AP)

The American Stock Exchange announced Monday it
was planning a tr
attempt to link U
markets during th
The exchange s
develop a busines
American and Far
joint ventures th
trading facility
``As global fi
1990s there will
foreign securitie
and Pacific rim c
Exchange Chairman
The business d
New York and Toky
markets.
The Amex is th

Today in History
Today is Friday, Aug. 12, the 225th day of
1988. There are 141 days left in the year. Today's
highlight in history: On Aug. 12, 1898, Hawaii was
formally annexed to the United States after
Congress passed a joint resolution. Hawaii was
granted territorial status in 1900, and became the
50th state of the union in 1959.
On this date:
In 1851, Isaac Singer was granted a patent on
his sewing machine.
In 1867, President Andrew Johnson sparked a
move to impeach him as he defied Congress by
suspending Secretary of War Edwin M. Stanton.
In 1898, the peace protocol ending the Spanish-American War was signed.
In 1915, 75 years ago, the novel ``Of Human

Passage Retrieval

When did Hawaii become a state?

AP890309-0014 6.000720546052219 on a computer network he ordered installed to provide security at last year's two national political conventions and to meet senators' state office staff members ``I've got to be a people person '' he said ``They get to know who the sergeant-at-arms is when they pick up the phone '' Serving as the Senate's chief
purchasing offic
Giugni said ``I
computer system

AP890501-0067 7.375156643863451 without comment let stand rulings from Pennsylvania that included Hawaii in a so-called class-action settlement of claims against the asbestos companies Hawaii officials said they were not given a proper
opportunity to remo
action court settle
school districts ar
from an asbestos cl
excused from the ge

FT924-10620 6.000720546052219 from Mr Reed's advertisements. However the Republicans have not always been on the outside looking in. Before statehood was achieved in 1959, they dominated what was a federal territory with power inherited from missionaries and plantation owners. But the legacy turned into a burden as their party came to be perceived as elitist Plantation labourers, and their children and grandchildren now working in hotels have opted for

Answer Extraction

When did Hawaii become a state?

AP900416-0049 17.0832
House from 1954 to **1959**, the year Hawaii became

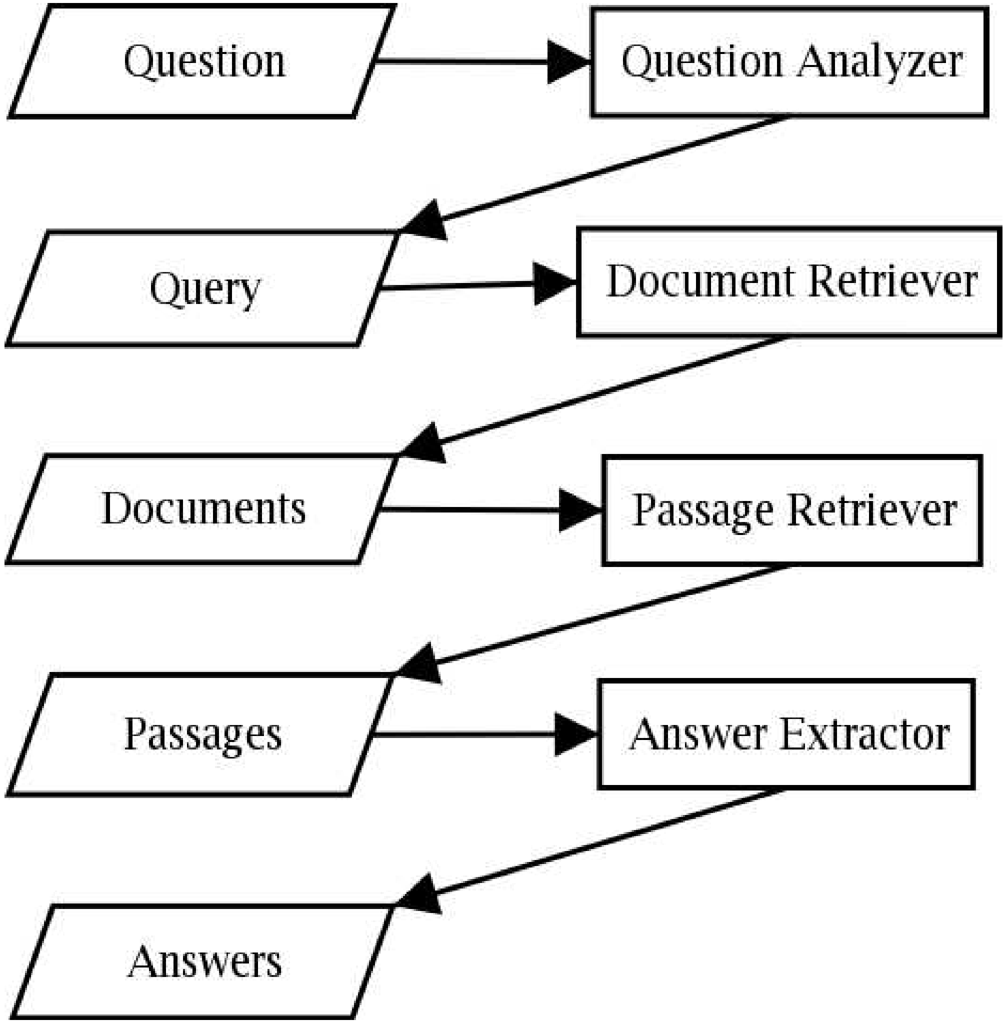
AP890417-0027 9.1485
Hawaii in 1974 became the first state in the

WSJ911010-0028 6.5864
on the Polynesian people of the 19th century.

FT924-10036 5.9691
Since becoming states in **1959**, however, no

SJMN91-06320033 4.8544
Since 1974, Hawaii has been the only state

Generic Question Answering Architecture



Our Experiments: Passage Retrieval

- Study a single component of question answering systems.
- Find out what passage retrieval techniques work.
- Make recommendations for improved question answering performance.

Why Passage Retrieval?

- Important module in many question answering systems.
- Not well studied before.
- Evidence that users prefer passage sized answers over exact answers because it gives context. (Lin *et al.*, *CHI 2003*)

Related Work

- Passage retrieval in the context of improving document retrieval performance.
  - Salton *et al.*, *SIGIR 1993*. Returned passages only if they were better than the document.
  - Callan, *SIGIR 1994*. Passage retrieval to improve the performance of document retrieval.
- No studies of passage retrieval for the question answering task (as far as we know).

Experimental Design

- Matrix experiment for question answering task.
- Three document retrievers.
  - Lucene
  - PRISE
  - oracle retriever
- Eight passage retrieval algorithms.
  - MITRE with stemming, MITRE without stemming, *bm25*, MultiText, IBM, SiteQ, Alicante, ISI.

Procedure

- Trained on the TREC 9 data set.
- Tested with TREC 10 data.
- Scored using percentage of unanswered questions and mean reciprocal rank (MRR).
- Computed both strict and lenient scores.
  - Lenient - Match one of the answer patterns provided by NIST.
  - Strict - Only relevant documents.

Mean Reciprocal Rank

- MRR (mean reciprocal rank)
  - Used at TREC QA tracks.
  - Invert the rank of the first correct answer, and average over all questions.
  - Between 0 and 1, higher is better.
  - Roughly correlated with percentage of unanswered questions.

Leveling the Playing Field

- Normalized passage lengths so every algorithm returned a 1000 byte answer.
  - Expanded or contracted the passage around the center point.
- Ran algorithms on the first 200 documents returned by the document retriever.

Document Retrievers

- Lucene
  - Boolean keyword search engine.
  - Typical of IR engines used by many TREC systems.
- PRISE
  - *bm25* term weighting.
  - Used the listing provided for TREC 10.
- oracle
  - Returns only documents that contain an answer.
  - Used the relevant document list from TREC 10.

Passage Retrieval Algorithms

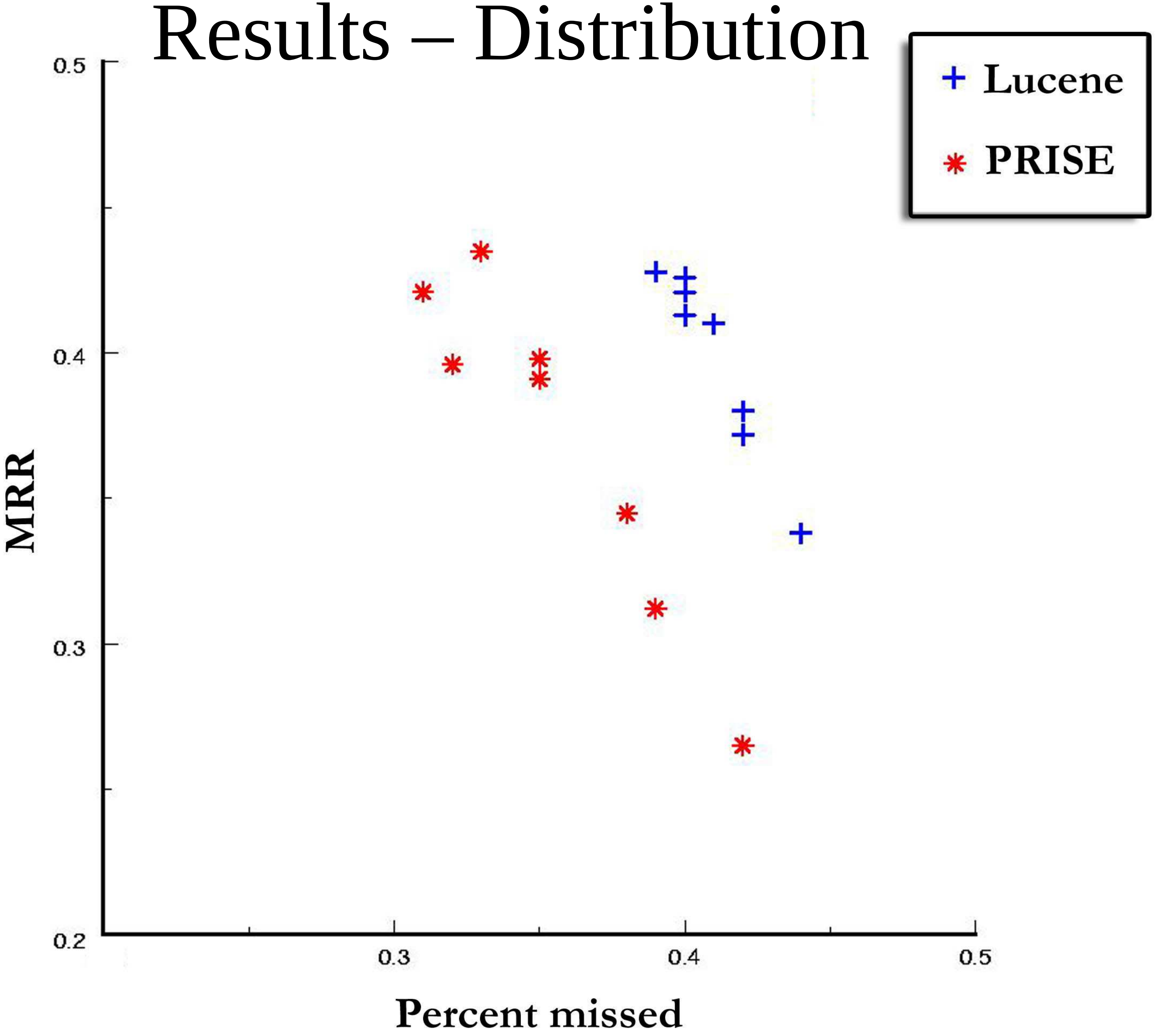
- Tokenizing
  - Sentence window
  - Word window
  - Query term window
- Weighting
  - *Constant*
  - *idf*
  - *bm25*
- Linguistic analysis
  - Synonyms (WordNet)
  - Stemming (WordNet, Porter)
- Tricks
  - Proper name match
  - Word co-location
  - Non length normalized cosine similarity

- Alicante (Llopis and Vicedo, *CLEF 2001*)
- *bm25* (Robertson *et al.*, *TREC 4*)
- IBM (Ittycheriah *et al.*, *TREC 9*)
- ISI (Hovy *et al.*, *TREC 10*)
- MITRE (Light *et al.*, *J. of Natural. Lang. Eng.*, *Special Issue on QA 2001*)
- MultiText (Clarke *et al.*, *TREC 9*)
- SiteQ (Lee *et al.*, *TREC 10*)

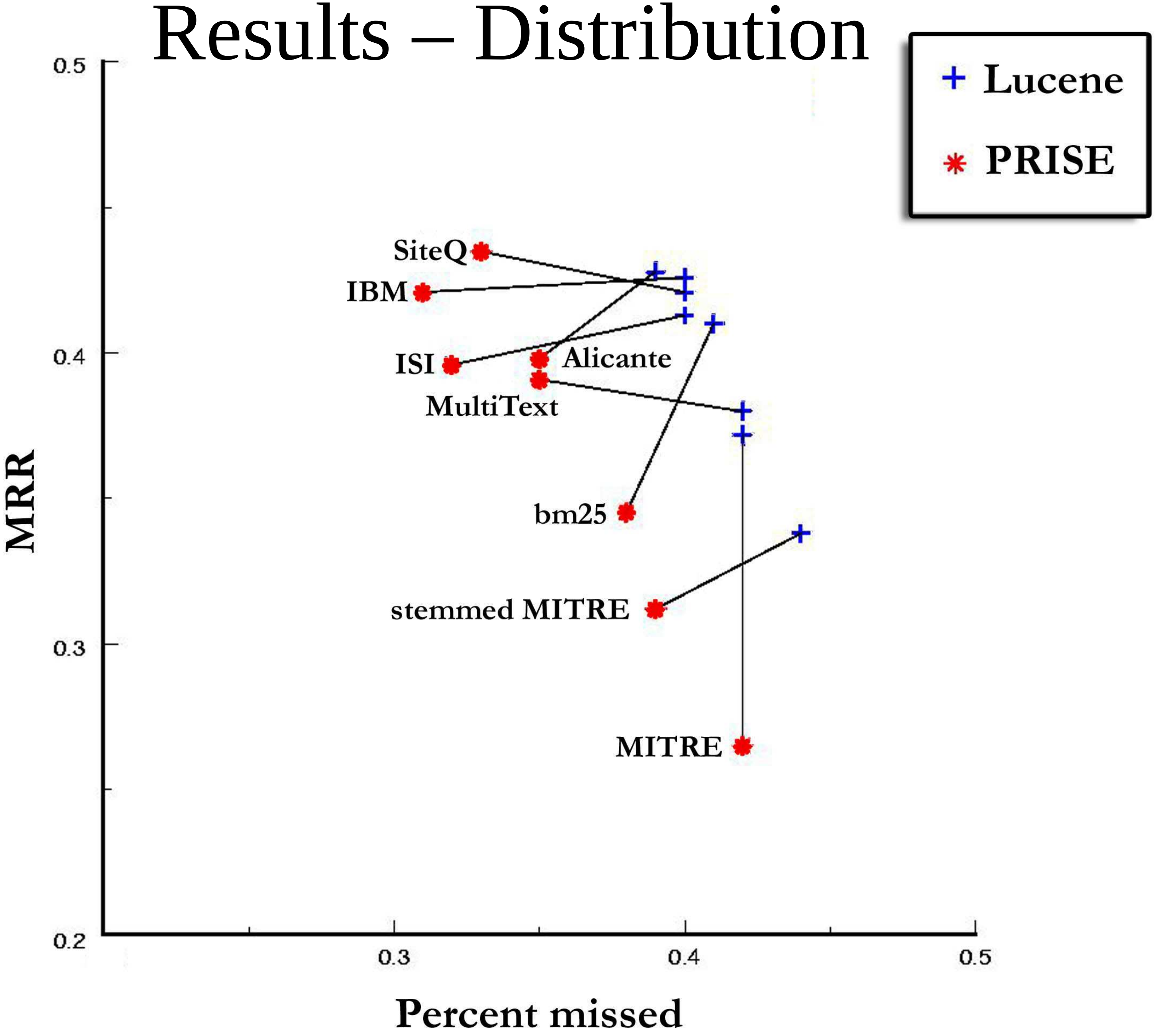
Algorithms Not Included

- InsightSoft (Soubbotin, *TREC 10*)
  - Cuts retrieved documents into passages around query terms, returning all passages from all retrieved documents.
  - Matching indicative patterns is fast.
- LCC (Harabagiu *et al*., *TREC 10*)
  - Retrieves passages containing keywords from the question based on the results of question analysis.
  - They did not describe their algorithm well enough for us to implement.

Results – Distribution

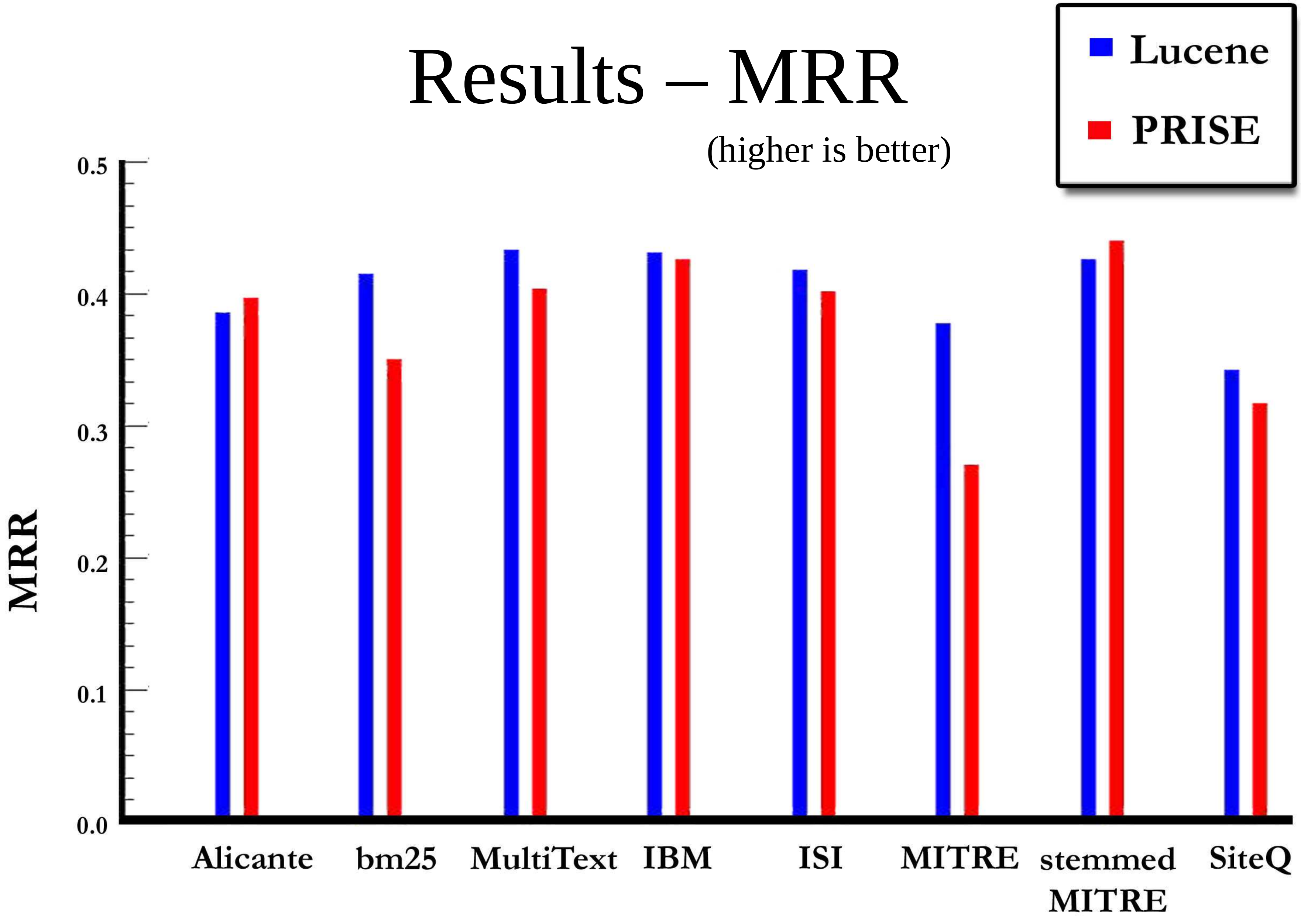


Results – Distribution

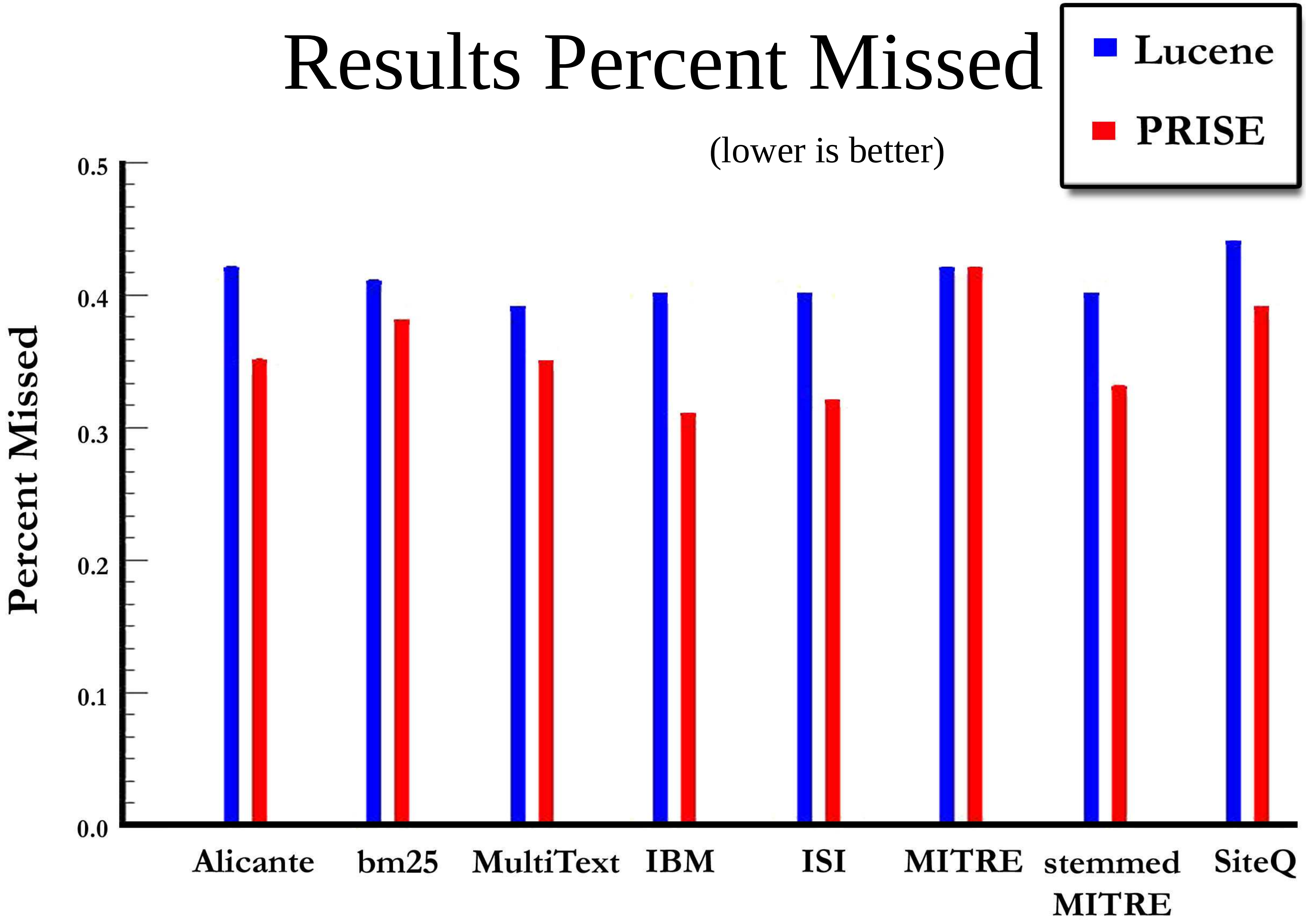


Results – MRR

(higher is better)



Results Percent Missed



Discussion - Boolean querying

- Lucene performed comparably to the PRISE document retriever.
- Boolean IR systems supply a reasonable set of documents for question answering.

Discussion – Passage Retrieval Algorithm Differences

- Lucene - differences among algorithms were not statistically significant.
  - Focus on document retrieval.
  - Focus on fundamentally different passage retrieval.
- PRISE - differences were significant.
  - Focus on improving passage retrieval and confidence ranking.
- oracle - differences were significant.
  - Passage retrieval is still an interesting problem.

Discussion – Density Based Scoring

- IBM, ISI, and SiteQ are statistically indistinguishable.
- All three give a non-linear boost to query terms that occur close together in the passage.
- IBM and ISI include proper case match and stemming.

Future Directions for Passage Retrieval

- Many missed definition questions.
  - Incorporate question type analysis to identify and handle them separately. (Prager *et al.*, *TREC 10*.)
- Others missed from ambiguous modification.
  - Example: What is the **highest dam** in the U.S.?
    - Extensive flooding was reported Sunday on the Chattahoochee River in Georgia as it neared its crest at Tailwater and George **Dam**, its **highest** level since 1929.
  - Recognize syntactic relations common to the question and the passage. (Katz and Lin, *EACL 2003*.)

Contributions

- Evaluated passage retrieval performance in isolation.
- Found that term density based passage retrieval algorithms work the best.
- Discovered that the relative performance of passage retrieval algorithms varies with the document retriever.

Results –- Lenient/MRR

Algorithm	Lucene	PRISE	oracle
Alicante	0.380	0.391	0.816
bm25	0.410	0.345	0.810
MultiText	0.428	0.398	0.845
IBM	0.426	0.421	0.851
ISI	0.413	0.396	0.852
MITRE	0.372	0.265	0.800
SiteQ	0.421	0.435	0.859
stemmed MITRE	0.338	0.312	0.762

Results – Lenient/Percent Missed

Algorithm	Lucene	PRISE	oracle
Alicante	41.80%	35.20%	9.03%
bm25	40.80%	38.00%	10.42%
MultiText	38.60%	34.80%	10.19%
IBM	39.60%	30.80%	7.18%
ISI	40.20%	32.20%	8.56%
MITRE	42.20%	42.00%	10.42%
SiteQ	40.20%	32.60%	7.41%
stemmed MITRE	44.20%	39.20%	14.58%

Results – Strict/MRR

Algorithm	Lucene	PRISE	oracle
Alicante	0.296	0.321	0.816
bm25	0.312	0.252	0.810
MultiText	0.354	0.325	0.845
IBM	0.326	0.331	0.851
ISI	0.329	0.287	0.852
MITRE	0.271	0.189	0.800
SiteQ	0.323	0.358	0.859
stemmed MITRE	0.250	0.242	0.762

Results – Strict/Percent Missed

Algorithm	Lucene	PRISE	oracle
Alicante	50.00%	42.60%	9.03%
bm25	48.80%	46.00%	10.42%
MultiText	46.40%	41.60%	10.19%
IBM	49.20%	39.60%	7.18%
ISI	48.80%	41.80%	8.56%
MITRE	49.40%	52.00%	10.42%
SiteQ	48.00%	40.40%	7.41%
stemmed MITRE	52.60%	58.60%	14.58%

Alicante

- Llopis and Vicedo, *CLEF 2001*.
- Six-sentence scoring window.
- Non-length normalized cosine similarity.
- Number of apperances of the term in the query and passage
- *idf* values of the terms.

Okapi *bm25*

- Robertson *et al.*, *TREC 4*.
- Basis of term weights:
  - Probability of appearing in relevant documents.
  - Probability of appearing in non relevant documents.
  - *tf* (term frequency in the document)
  - *idf* (inverse term frequency in the corpus)

IBM

- Ittycheriah *et al.*, *TREC 9*.
- Weighted sum of various distance measures.
  - Matching words –- Sums the *idf* of query terms that appear in the passage.
  - Thesaurus match - Sums the *idf* of query terms whose WordNet synonyms appear in the passage.
  - Mis-match words - Sums the *idf* of query terms that do not appear in the passage.
  - Dispersion - Counts the number of words in the passage between matching query terms.
  - Cluster words - Counts the number of words that occur adjacently in both the question and the passage.

ISI

- Hovy *et al.*, *TREC 10*.
- Weighted sum of various features:
  - Exact match of proper names
  - Match of query terms
  - Match of stemmed terms
- We ignored answer extraction correction term.

MITRE

- Light *et al.*, *J. of Natural. Lang. Eng.*, *Special Issue on QA 2001*.
- Baseline algorithm
  - Tokanizes into 1-sentence windows
  - Counts the number of query terms that appear in the sentence.
- Stemming and non stemming versions.

MultiText

- Clarke *et al.*, *TREC 9*.
- Favors short passages with many query terms.
- *idf* term weighting.
- Tokenizes on query terms.

SiteQ

- Lee *et al.*, *TREC 10*.
- 3 sentence passage window.
- Density based weighting.
- *idf* weight instead of part of speech weights.