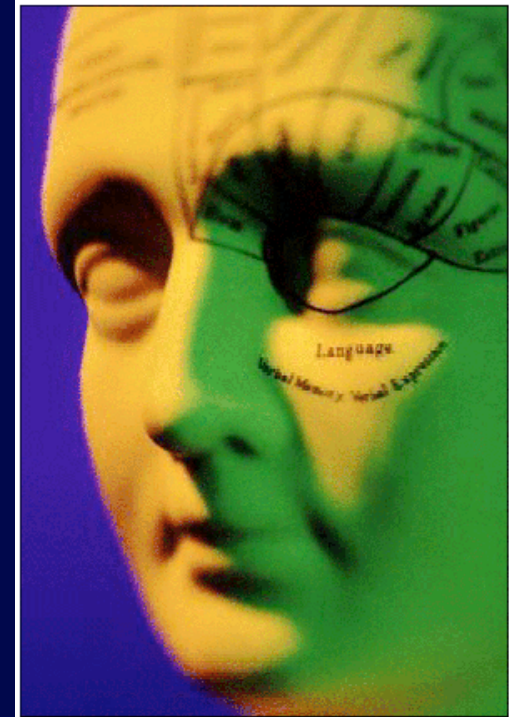# Question Answering Engineering

Zhiping Zheng

Computational Linguistics Department
Saarland University, Germany
**zheng@coli.uni-sb.de**
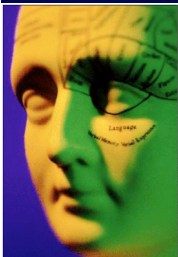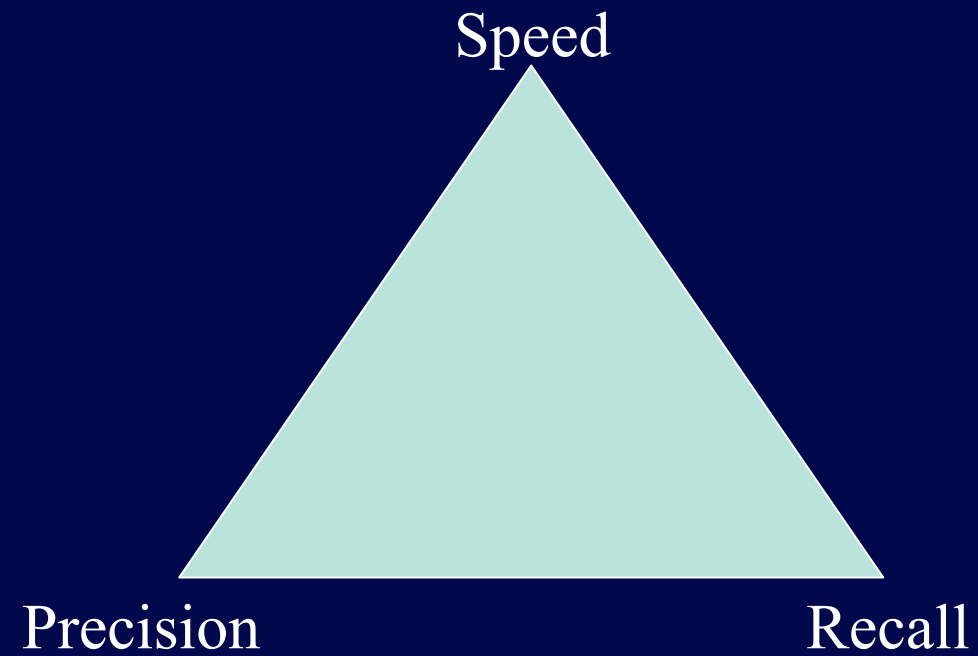
Saarbrücken, Germany. Dec 13, 2004

Information on the Web

The Web is getting bigger and bigger

- You have more and more chances to find the information you need on the Web
- Search engine is the tool for all Web users
- More and more people (companies, organizations, etc.) use Web to publish the information the want to make public
- More and more search engines (tools to help Web users to retrieve information on the Web) are available for different purpose
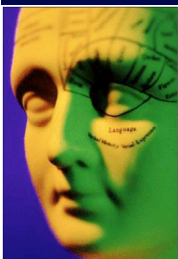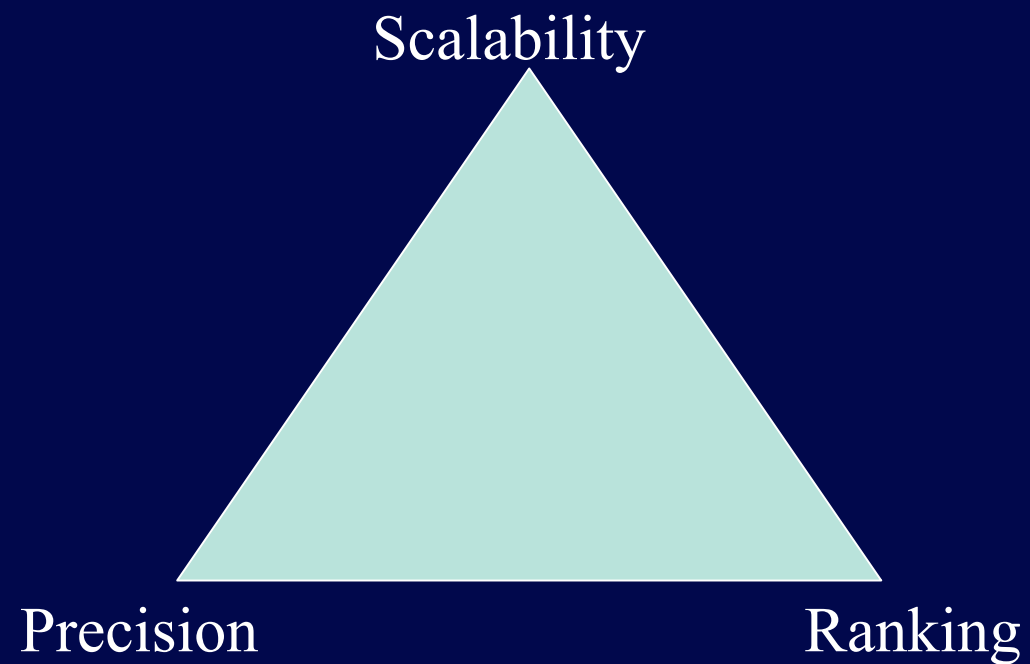
Saarbrücken, Germany. Dec 13, 2004

Search Engine Triangle

Speed

Precision

Recall
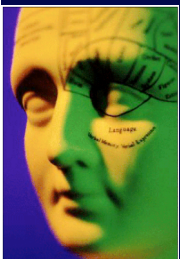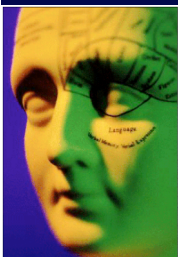
Saarbrücken, Germany. Dec 13, 2004

New Search Engine Triangle

Scalability

Precision   Ranking

Scalability

- Works for both large size and small size data?
- Works for large amount of traffics?
- Works in distributional/parallel environment?
- Language independent?
- Is there a reasonable caching mechanism?
- Know what to index?

Question Answering: Introduction

○ **What is question answering?**

- User has a question, system has to return an answer (Buchholz-Tilburg Univ.)

- A computer-based activity that involves searching large quantities of text and understanding both questions and textual passages to the degree necessary to recommend a text fragment as an answer to a question. (Univ. MASS)

- Provides direct answers to user questions by consulting its knowledge base. (Mulder - Univ. of Washington)

- Special information retrieval using natural language processing and artificial intelligence techniques.
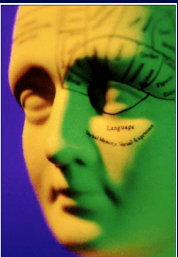
Saarbrücken, Germany. Dec 13, 2004
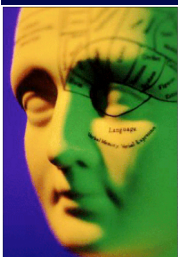
A question answering system is ...

- NOT a search engine
- NOT a chatter bot
- NOT a dialogue system
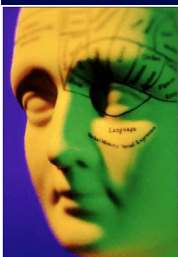- NEITHER an expert system NOR a knowledge base
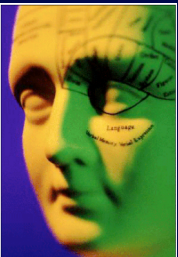
Why question answering?

- Question answering provides intuitive information access

- People need just the right information

More on Question Answering

○ **What is Web-based question answering?**

- Accessible from Web
- Uses Web as its knowledge base

○ **What is open-domain question answering?**

- Technically domain insensitive
- Practically able to answer questions in different domains

Saarbrücken, Germany. Dec 13, 2004

- **Factual questions**
  - Who wrote King Lear?
- **Simple reasoning**
  - Why is the sky blue?
- **Some other questions**
  - Why is New York City called The Big Apple?
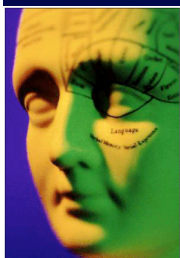
What can be answered by a QA System?

- The answers come from people's knowledge
- The answers should be retrievable
  - How many hearts does an octopus have?
  - How many hearts does a peacock have?
- The question should be understandable by the system
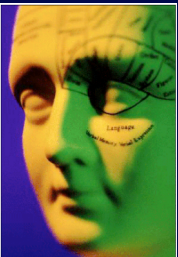
Architecture and Basic Components of QA Systems

○ **Basic Components**

- Question processing
- Search
- Answer selection and ranking
- Answer generation
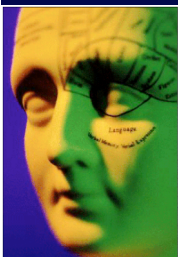
○ **Requirements of Components**

- Effectiveness
- Scalability
- System load

Output Format Comparison

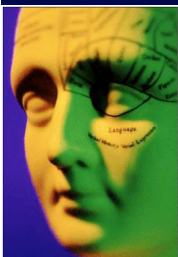| QA System	Output
AnswerBus	Sentences
AskJeeves	Documents
IONAUT	Passages
LCC	Sentences
Mulder	Extracted answers
QuASM	Document blocks
START	Mixture
Webclopedia	Sentences

Saarbrücken, Germany. Dec 13, 2004

- One-word exact answer?
  - Do you believe that is a correct answer?
- Detailed answer in a paragraph?
  - You still need to explore the answer
- A sentence with contextual information?
  - The choice of AnswerBus and some other QA systems
- Source links?
  - What is the difference from search engines?
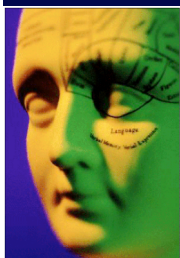
Web as a QA knowledge base

- Answers: treasures hidden in the Web
- Search engines: pointers to the treasures
- Effective algorithm: Tools for exploring the treasurers
- Pre-tagging: impossible except you also make the search engine

Saarbrücken, Germany. Dec 13, 2004

QA Research Systems on the Web

- AnswerBus (Zhiping Zheng)
- IONAUT (Steve Abney and Michael Collins)
- Lamp (National Singapore University)
- LCC (Language Computer Corp.)
- Mulder (U Washington)
- QuASM (U MASS)
- START (MIT/AI Lab)
- Webclopedia (USC)

Saarbrücken, Germany. Dec 13, 2004
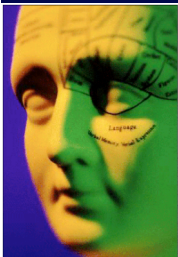
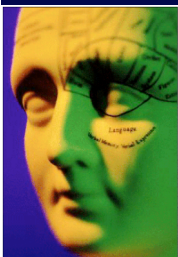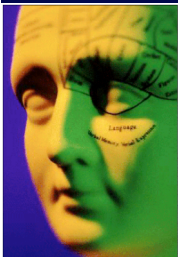# Commercial QA Systems

- AskJeeves
- FlexAnswer
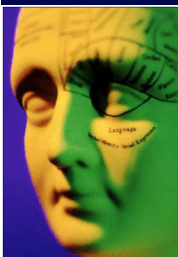- InfoJukeBox
- iPhrase
- Primus (formerly AnswerLogic)

Combination of domain specific system and open-domain system

- A domain-specific QA system is only powerful for questions in its specific domain. START can only answer 29 out of 200 TREC-8 questions. For this 29 questions, the correct answer rate is 100%. If we consider all 200 questions, the correct answer rate is 14.5%.

- A domain-independent QA system can answer questions in any domains but it is not easy for such a system to answer expert-level questions in any domain. AnswerBus can give 70.5% correct answer rate to 200 TREC-8 questions overall but it is clearly not as good as START for those 29 questions answered by START.

- $29 \times 100\% + (200-29) \times 70.5\% = 29+120.6 = 149.6$

Saarbrücken, Germany. Dec 13, 2004

TREC Evaluation Metrics:

- Select a big corpus as the knowledge base
- Use same question set for all systems
- Mean Reciprocal Rank (MRR)
  - Inverse of rank at which first correct answer was found
  - MRR = average over all questions
  - Judgments: correct, unsupported, incorrect

More on TREC Measurement
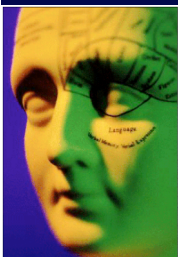
○ **What does TREC measurement:**

- Precision combined with ranking

○ **What else should be measured**

- Speed of response
- System requirements
- Corpus requirements
- Other components requirements
- User input requirements
- Interface

Saarbrücken, Germany. Dec 13, 2004
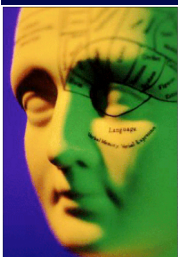
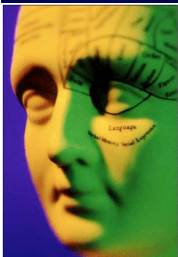- Real-time evaluation
- Task oriented assessment
- User feedback
- System popularity
- Words from experts

System Popularity

- Position in search engines
- Number of pages containing special terms
- Links to the system
- Alexa traffic rank
- User Groups and Open Directory
- Web testimonials

Saarbrücken, Germany. Dec 13, 2004

Question Answering Triangle

Speed

Precision

Temporality
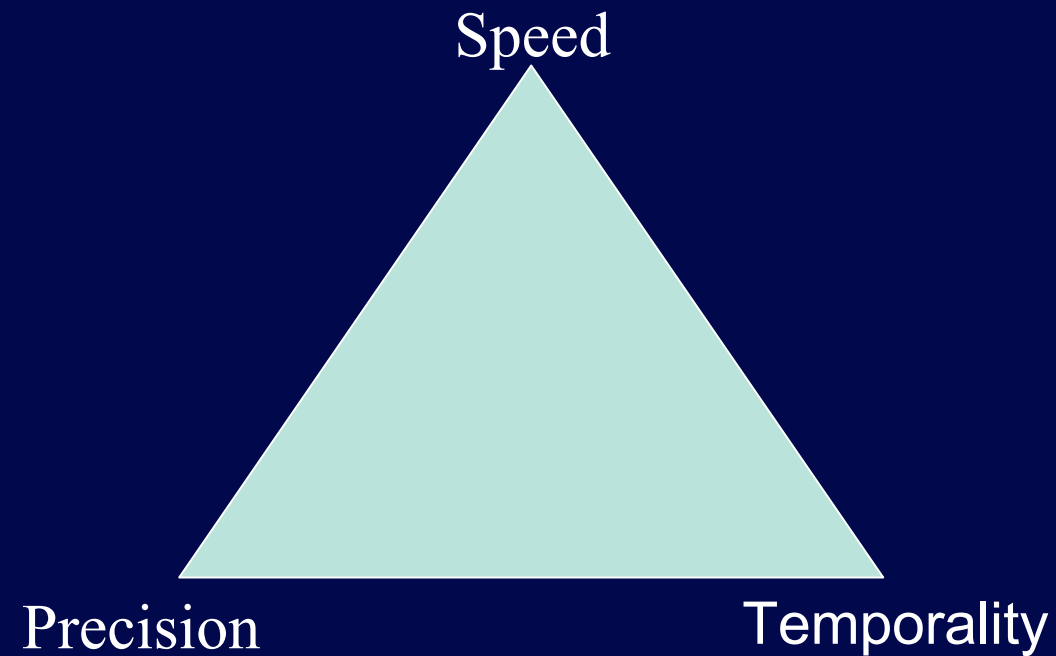
A Question Answering Survey

- Questionare:
  - http://survey-qa.dfki.de/survey-qa-en.html

- Results:
  - http://survey-qa.dfki.de/answers.html

Saarbrücken, Germany. Dec 13, 2004

Corpus as a knowledge base

- Size of a corpus required by a QA system
- Dynamic vs. static
- Domain specific vs. open-domain
- Local vs. distributional
- Structure of a corpus (domain specific)
- Layer structure and tree structure
- Automated construction vs. human picking

QA Strategy and methodology

- Language ambiguities and semantic principle
- AnswerBus' Question-sentence matching formula

$$q \geq \left\lfloor \sqrt{Q-1} \right\rfloor + 1$$
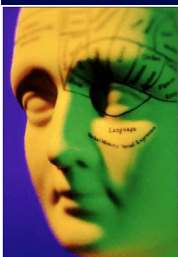
- q is the number of matching words in the sentence.
  Q is the total number of matching words in the question.

- Uncertainty principle
- Baseline principle
- Bottle neck and trade-off
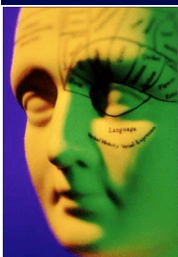
Other issues:

- Top down or bottom up?
- Question categorization?
- Answer type
- Topic detection
- Temporal issue
- Contextual information
- Multilinguality

Question Answering Environment:

- Operating system
- Corpus structure and file system
- Caching mechanism
- Programming language
- File exchange
- Parallel / pseudo parallel algorithm
- Load test

Saarbrücken, Germany. Dec 13, 2004
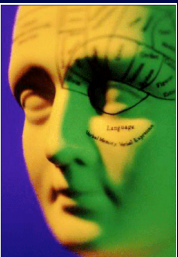
QA Load Test

- Running the system on a single machine
- Disable caching system
- Send a question to the server every second and every 3 seconds
- Read the system load image

Saarbrücken, Germany. Dec 13, 2004

System Load Image
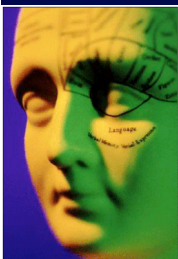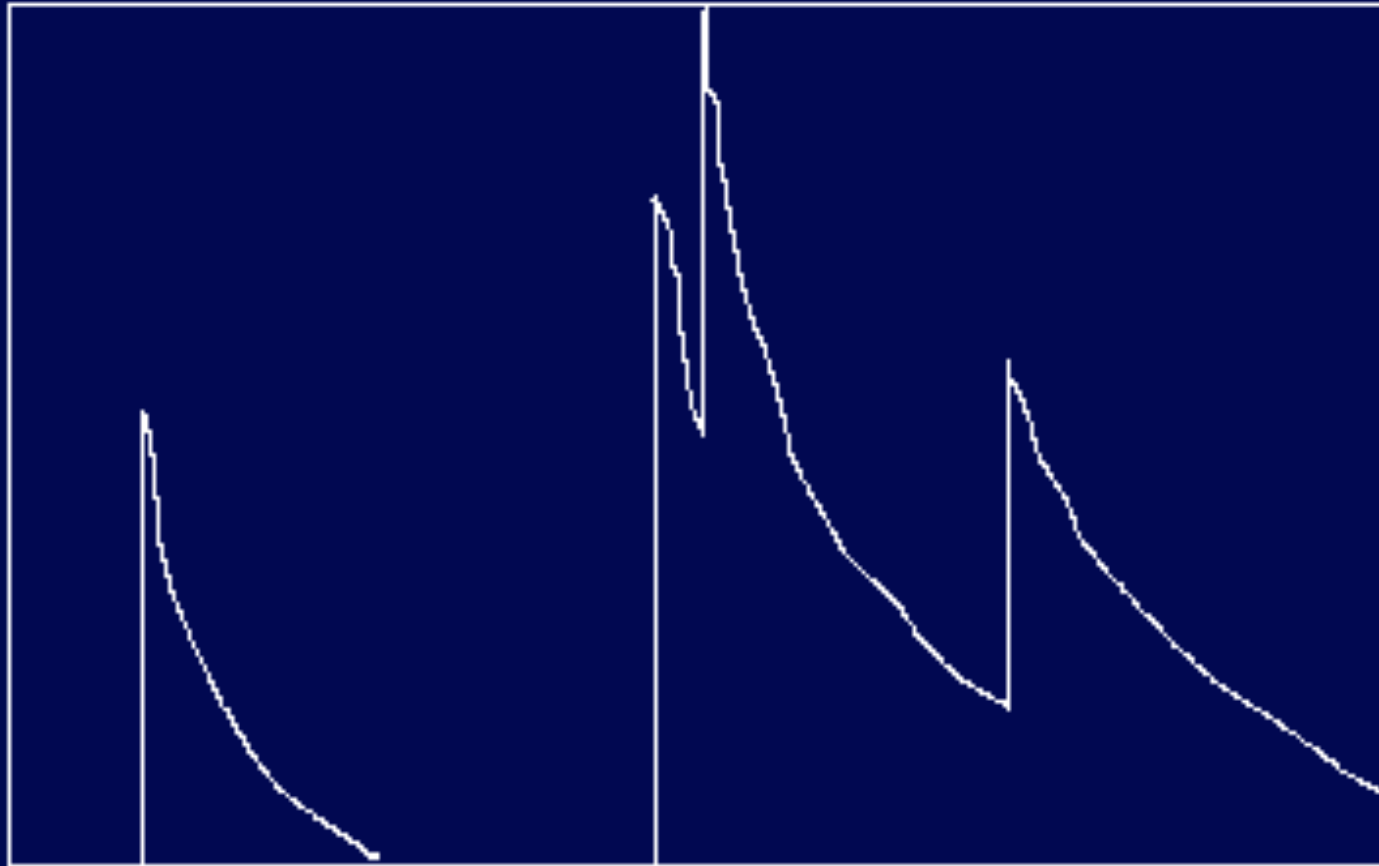
NLP Enhancement

- Deep or shallow
- Sentence segmentation
- Use of related words
- Word form modification
- Spell check and spell suggestion
- QA dictionary
- Cross language search

Saarbrücken, Germany. Dec 13, 2004

IR and IE Components

- Key word selection and document search
- Sentence level search and indexing
- QA specific indexing
- Intelligent crawler
- Gazetteers
- Human knowledge base
- Dynamic named-entity extraction
- Intelligent search vs. question answering

Does caching system work?

- Search strings leading to AnswerBus
  - (August 1 – 9, 2003)

1	989	1.99%	how many feet in a mile
2	635	1.28%	first king of england
3	597	1.20%	how many square feet in an acre
4	484	0.97%	acre square feet
5	467	0.94%	oldest us president
6	374	0.75%	oldest president
7	367	0.74%	kilos to pounds
8	325	0.65%	feet in a mile
9	300	0.60%	calories banana
10	277	0.56%	square feet per acre

- ## Question log of AnswerBus
  - – + What is North America's favorite snack food?
  - – + What is the most popular sport in Japan?
  - – + How do I copy Playstation Games?
  - – + what is the chemical symbol for ammonia?
  - – + Why does our skin get thin as we age?
  - – 10 Philadelphia Fairmount?
  - – + What is the most popular sport in the world?
  - – + how can i repair scratched monitor?
  - – + What color is indigo?
  - – 5 "J" Movies?
  - – + When did Elvis Presley die?
  - – 5 what do the initials U.N.E.S.C.O. stand for?
  - – 1 What is the abbreviation for Texas?
  - – + How long to cook a steak?
  - – + Is Croatia a safe place?
  - – + where can i find ranma wallpaper?

Saarbrücken, Germany. Dec 13, 2004
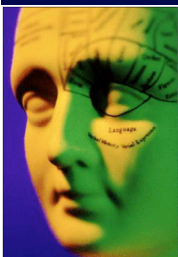
What is AnswerBus?

- AnswerBus (http://www.answerbus.com/) is an open-domain natural language question answering system (QA) and intelligent search engine.

- It consists of a Web system that uses the entire Internet as its knowledge base, and a local system that can be implemented on any collection of information, or corpus.

- One example of local implementation is the AnswerBus News Engine (http://www.answerbus.com/news/), which was based on a collection of 700,000 CNN news pages.

Original AnswerBus QA System
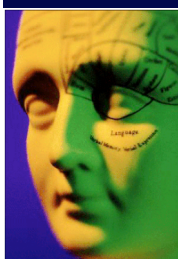
○ Running on a shared Linux server in University of Michigan, U.S.

○ Using Google and other four commercial search engines

○ Accept six languages

○ Correctly answered 70.5% TREC-8 questions

○ Average response time : 7 seconds

○ Thousands of visits every day

○ Being used as a sample system by some NLP courses

Saarbrücken, Germany. Dec 13, 2004
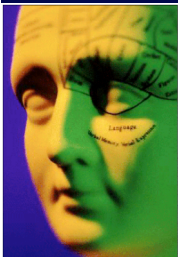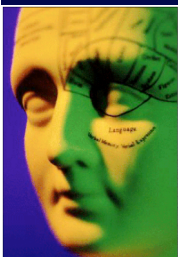
AnswerBus: Processing Diagram
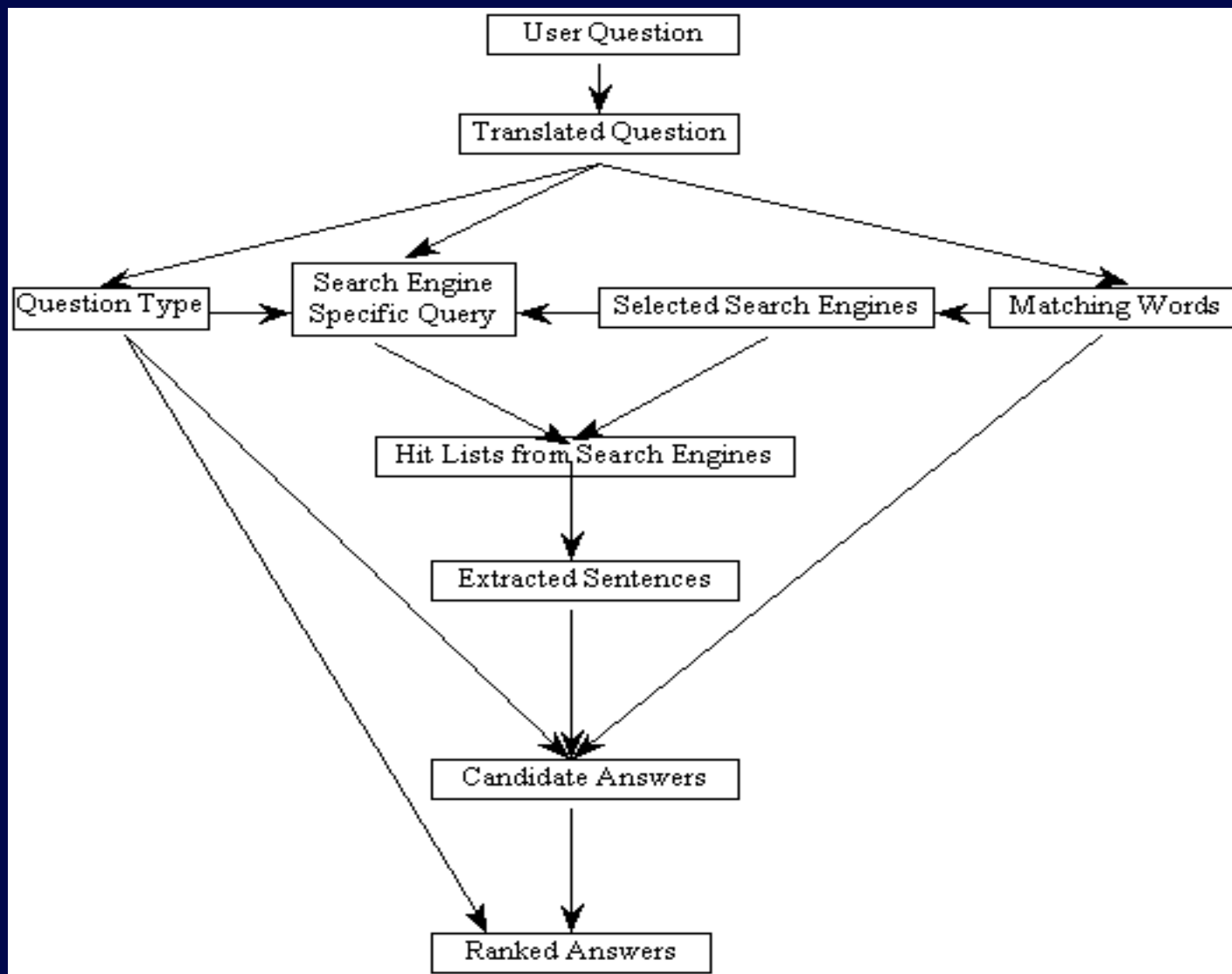
Differences to Other Question Answering System

	TOP 5	TOP 1	NIST SCORE	Tmean (seconds)	Lmean (bytes)
AnswerBus	141	120	0.642	7.20	141
IONAUT				12.51	1312
LCC	97	75	0.417	44.24	178
QuASM	13	7	0.045	20.72	1766
START	29	29	0.145	9.84	

Saarbrücken, Germany. Dec 13, 2004

○**Use of Search Engines**

- 5 basic search engines:
  - Google, Yahoo, WiseNut, AltaVista
  - and Yahoo News
  - Maybe more: Northern Light, Medline Plus

○**Search Engine Selection**

○**Search Engine Specific Query**

- Functional words deletion
- Special words deletion
- Word form modification

Saarbrücken, Germany. Dec 13, 2004

AnswerBus: Plain Text Sentence Segmentation

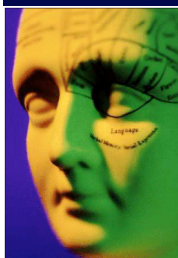○ **Ending with a dot (.) in most case**

○ **Sometimes ending with question mark (?), exclamation mark (!), carriage return, or something else.**

○ **Many exceptions:**

  ▪ **"." in some words: U.S., pp., Mr., etc.**

○ **General methods for segmentation in plain texts:**

  ▪ **Regular expression**

  ▪ **perl**

  ▪ **Pattern matching**

  ▪ **Not working for some cases**

Saarbrücken, Germany. Dec 13, 2004

AnswerBus: HTML Sentence Segmentation

○ **Having all features with plain text**
○ **Tags related to sentence boundaries:**
  ▪ **<p>, <br>, <h1>, etc**
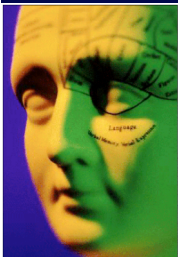○ **HTTP head, HTML comments, JavaScript, font tags, etc.**
○ **XML entities:**
  ▪ ** > ü etc.**
○ **HTML table and other format tags**
○ **More exceptions: "." in URLs or email addresses etc.**
○ **General methods for sentence segmentation in HTML document**
  ▪ **Delete all tags then process as plain text**
○ **Goals**
  ▪ **A sentence segmenter for both HTML and plain text**
  ▪ **Running at a high speed**

Saarbrücken, Germany. Dec 13, 2004

Candidate Answer Extraction & Ranking

- Question type and QA specific dictionary
  - Question type
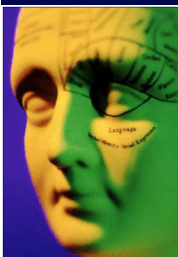 - "How far" is different from "how close"
- Dynamic basic named entities extraction
  - Pre-tagging and run-time tagging
  - Only very few named entities need to be extracted for a specific user question
- Coreference resolution
- Redundancy
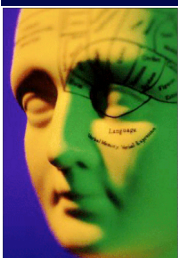
- who is arthur kravitz?
- What is the first Gothic novel?
- Burmese refugees in Thailand?
- What percentage of 16 year olds are still virgins?
- Which famous building was built by Shih Huang Ti ?
- What is Melanin?
- why was the iditarod race started?
- what are MARC 008 tags?
- Who was the grand old Duke of York?
- Philip roth biography?

Saarbrücken, Germany. Dec 13, 2004

AnswerBus' caching system

○ **AnswerBus' caching system**
  - First layer: question layer
  - Second layer: Web documents

○ **Cache management**
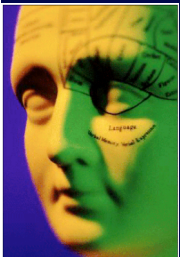
AnswerBus CNN News Engine

http://www.answerbus.com/news/

- ❍ Over 700,000 HTML documents indexed locally
- ❍ The news data are segmented into sentences
- ❍ 1-4 seconds to answer one question
- ❍ Update the index database every 2 hours
  40 minutes if Linux cluster is working
- ❍ Running on a single Linux machine independently
- ❍ A by-product: CNN Web site search engine
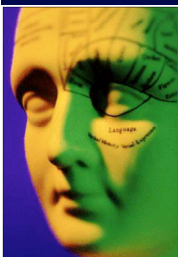
Saarbrücken, Germany. Dec 13, 2004

New Features

○ Largest text data base for QA task

○ Sentence level indexing and searching
- Fast
- Can be updated partially and logically

○ Embedded search engine

○ Scalable to very large scale

○ Dynamic information matching and extraction

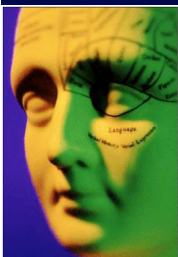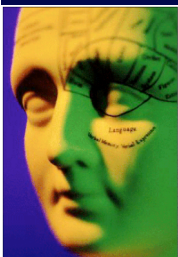# Differences to Other Search Engines

○ **It is a search engine, a search engine with natural language interfaces.**

○ **As compared to regular search engines, AnswerBus provides a more natural interface that leads users directly to the very needed answers, thus can significantly shorten users' navigation path and ease the information retrieval effort.**

○ **Its local system can be implemented in corporations and organizations, for example, on an Intranet site, to help employees easily locate job related information and enhance productivity. Thus, it can be also used as a local search engine.**

Saarbrücken, Germany. Dec 13, 2004

Search engine as QA system

- Fully n-gram and sentence level indexing
- Timely index updating
- Corpus size controlling (< 200G)
- Powerful input understanding
- Embedded expert systems
- Can be running on a single machine and also a distributional system

Saarbrücken, Germany. Dec 13, 2004

AnswerBus Achievements

Since its debut in 2002, AnswerBus has achieved favorable reputation in both academia and the Internet community.

- It was used as an example NLP application in different universities, including Edinburgh, Charles, Masaryk, MIT, Stanford, Columbia, and Pittsburgh.
- The traffic to the website has been continuously increasing.
- Research papers about AnswerBus were presented to the most prestigious conferences and conventions in the field of computational linguistics and language technology, including HLT, WWW and ACL.
- It was cited by numerous academic research papers and thousands of Web pages.
- It was highly regarded by experts and end users. Reviews and comments about AnswerBus can be found at http://www.answerbus.com/testimonials/ .
- Many research groups have expressed interests in using AnswerBus in both academic research and/or industrial applications.

Saarbrücken, Germany. Dec 13, 2004
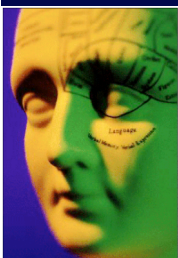
Future Research Work

- Rigorous Evaluation Method
- Domain Specific QA System
- Multilingual
- More NLP
- Answer Generation
- Computer Telephony
- Dialogue understanding
- Speaking interface of question answering
- Memory and logical reasoning

Saarbrücken, Germany. Dec 13, 2004