

# Corpus-Based Question-Answering

Introduction to Information Retrieval

Christof Monz and Maarten de Rijke

Spring 2002

# Outline

# Outline

- What's corpus-based Q&A

# Outline

- What's corpus-based Q&A
- Q&A and TREC

# Outline

- What's corpus-based Q&A
- Q&A and TREC
- Current approaches to Q&A

# Outline

- What's corpus-based Q&A
- Q&A and TREC
- Current approaches to Q&A
- The role of NLP in IR and Q&A

# Outline

- What's corpus-based Q&A
- Q&A and TREC
- Current approaches to Q&A
- The role of NLP in IR and Q&A
- The role of semantics in Q&A

# Outline

- What's corpus-based Q&A
- Q&A and TREC
- Current approaches to Q&A
- The role of NLP in IR and Q&A
- The role of semantics in Q&A
- Our working plan

# Outline

- What's corpus-based Q&A
- Q&A and TREC
- Current approaches to Q&A
- The role of NLP in IR and Q&A
- The role of semantics in Q&A
- Our working plan
- Conclusions

# Traditional Question-Answering

# Traditional Question-Answering

- Traditional approaches assume the data to be captured in a structured format, such as databases

# Traditional Question-Answering

- Traditional approaches assume the data to be captured in a structured format, such as databases
- Questions are translated into a query language (SQL)

# Traditional Question-Answering

- Traditional approaches assume the data to be captured in a structured format, such as databases
- Questions are translated into a query language (SQL)
- Pros:

# Traditional Question-Answering

- Traditional approaches assume the data to be captured in a structured format, such as databases
- Questions are translated into a query language (SQL)
- Pros: + Good access control

# Traditional Question-Answering

- Traditional approaches assume the data to be captured in a structured format, such as databases
- Questions are translated into a query language (SQL)
- Pros: + Good access control  
+ efficient run time behavior

# Traditional Question-Answering

- Traditional approaches assume the data to be captured in a structured format, such as databases
- Questions are translated into a query language (SQL)
- Pros:
  - + Good access control
  - + efficient run time behavior
- Cons:

# Traditional Question-Answering

- Traditional approaches assume the data to be captured in a structured format, such as databases
- Questions are translated into a query language (SQL)
- Pros: + Good access control  
+ efficient run time behavior
- Cons: – Generating structured databases is expensive

# Traditional Question-Answering

- Traditional approaches assume the data to be captured in a structured format, such as databases
- Questions are translated into a query language (SQL)
- Pros:
  - + Good access control
  - + efficient run time behavior
- Cons:
  - Generating structured databases is expensive
  - translating NL to SQL has to be very reliable (problematic for open domains)

# Q&A and Information Retrieval

# Q&A and Information Retrieval

- Corpus-based Q&A grew out of document retrieval

# Q&A and Information Retrieval

- Corpus-based Q&A grew out of document retrieval
- Document retrieval task:

# Q&A and Information Retrieval

- Corpus-based Q&A grew out of document retrieval
- Document retrieval task: Given a collection of documents (articles, web pages, . . . ) and a query, identify documents that are relevant to the query

# Q&A and Information Retrieval

- Corpus-based Q&A grew out of document retrieval
- Document retrieval task: Given a collection of documents (articles, web pages, . . . ) and a query, identify documents that are relevant to the query
- Most popular systems: Internet search engines (Google, AltaVista, Lycos, . . . )

# Q&A and Information Retrieval

- Corpus-based Q&A grew out of document retrieval
- Document retrieval task: Given a collection of documents (articles, web pages, . . . ) and a query, identify documents that are relevant to the query
- Most popular systems: Internet search engines (Google, AltaVista, Lycos, . . . )
- A document is not an answer to a question

# Q&A and Information Retrieval

- Corpus-based Q&A grew out of document retrieval
- Document retrieval task: Given a collection of documents (articles, web pages, . . . ) and a query, identify documents that are relevant to the query
- Most popular systems: Internet search engines (Google, AltaVista, Lycos, . . . )
- A document is not an answer to a question
- Relevant documents often don't provide an answer

# Answer Selection

# Answer Selection

- Given a question and a list of (presumably) relevant documents

# Answer Selection

- Given a question and a list of (presumably) relevant documents
- Analyze those documents to find an answer

# Answer Selection

- Given a question and a list of (presumably) relevant documents
- Analyze those documents to find an answer
- Return the answer to the user

# Answer Selection

- Given a question and a list of (presumably) relevant documents
- Analyze those documents to find an answer
- Return the answer to the user
- The analysis of those documents involves several shallow NLP techniques

# The Role of Retrieval in Q&A

# The Role of Retrieval in Q&A

- Why rely on imperfect document retrieval to identify relevant documents?

# The Role of Retrieval in Q&A

- Why rely on imperfect document retrieval to identify relevant documents? Why not analyze all documents in the collection?

# The Role of Retrieval in Q&A

- Why rely on imperfect document retrieval to identify relevant documents? Why not analyze all documents in the collection?
- Standard collections for Q&A contain around 1,000,000 documents

# The Role of Retrieval in Q&A

- Why rely on imperfect document retrieval to identify relevant documents? Why not analyze all documents in the collection?
- Standard collections for Q&A contain around 1,000,000 documents
- Analyzing all documents will take quite some time

# The Role of Retrieval in Q&A

- Why rely on imperfect document retrieval to identify relevant documents? Why not analyze all documents in the collection?
- Standard collections for Q&A contain around 1,000,000 documents
- Analyzing all documents will take quite some time
- Using IR as a pre-fetch allows for more focused analysis

# Is there a Place for NLP in Q&A

# Is there a Place for NLP in Q&A

- Document retrieval doesn't make any use of linguistic structure

# Is there a Place for NLP in Q&A

- Document retrieval doesn't make any use of linguistic structure
- Documents are represented as *bags-of-words*

# Is there a Place for NLP in Q&A

- Document retrieval doesn't make any use of linguistic structure
- Documents are represented as *bags-of-words*
- Although bags-of-words are sufficient for retrieval, Q&A requires a 'deeper' analysis

# Is there a Place for NLP in Q&A

# Is there a Place for NLP in Q&A

- Consider a question such as  
*Who killed Lee Harvey Oswald?*
- Document retrieval doesn't make a difference between:

# Is there a Place for NLP in Q&A

- Consider a question such as  
*Who killed Lee Harvey Oswald?*
- Document retrieval doesn't make a difference between:
  - ▶ *Jack Ruby killed Lee Harvey Oswald*

# Is there a Place for NLP in Q&A

- Consider a question such as  
*Who killed Lee Harvey Oswald?*
- Document retrieval doesn't make a difference between:
  - ▶ *Jack Ruby killed Lee Harvey Oswald*  
{Harvey, Jack, killed, Lee, Oswald, Ruby}

# Is there a Place for NLP in Q&A

- Consider a question such as  
*Who killed Lee Harvey Oswald?*
- Document retrieval doesn't make a difference between:
  - ▶ *Jack Ruby killed Lee Harvey Oswald*  
{Harvey, Jack, killed, Lee, Oswald, Ruby}
  - ▶ *Lee Harvey Oswald killed John F. Kennedy*

# Is there a Place for NLP in Q&A

- Consider a question such as  
*Who killed Lee Harvey Oswald?*
- Document retrieval doesn't make a difference between:
  - ▶ *Jack Ruby killed Lee Harvey Oswald*  
{Harvey, Jack, killed, Lee, Oswald, Ruby}
  - ▶ *Lee Harvey Oswald killed John F. Kennedy*  
{F., Harvey, John, Kennedy, killed, Lee, Oswald}

# Is there a Place for NLP in Q&A

- Consider a question such as  
*Who killed Lee Harvey Oswald?*
- Document retrieval doesn't make a difference between:
  - ▶ *Jack Ruby killed Lee Harvey Oswald*  
{Harvey, Jack, killed, Lee, Oswald, Ruby}
  - ▶ *Lee Harvey Oswald killed John F. Kennedy*  
{F., Harvey, John, Kennedy, killed, Lee, Oswald}
- Obviously, structural information is needed

# Measuring Quality

# Measuring Quality

- Assessing the quality of a Q&A systems requires objective evaluation

# Measuring Quality

- Assessing the quality of a Q&A systems requires objective evaluation
- For a set of questions one has to check whether the system returns the correct answer

# Measuring Quality

- Assessing the quality of a Q&A systems requires objective evaluation
- For a set of questions one has to check whether the system returns the correct answer
- This is a laborious process

# Outline

- What's corpus-based Q&A ✓
- Q&A and TREC
- Current approaches to Q&A
- The role of NLP in IR and Q&A
- The role of semantics in Q&A
- Our working plan
- Conclusions

# Outline

- What's corpus-based Q&A ✓
- Q&A and TREC
- Current approaches to Q&A
- The role of NLP in IR and Q&A
- The role of semantics in Q&A
- Our working plan
- Conclusions

# Text REtrieval Conferences (TREC)

# Text REtrieval Conferences (TREC)

- TREC is an annual conference/competition devoted to information retrieval related tasks

# Text REtrieval Conferences (TREC)

- TREC is an annual conference/competition devoted to information retrieval related tasks
- First TREC was held in 1992, since TREC-8 (1999) Q&A is one the tracks

# Text REtrieval Conferences (TREC)

- TREC is an annual conference/competition devoted to information retrieval related tasks
- First TREC was held in 1992, since TREC-8 (1999) Q&A is one the tracks
- TREC is sponsored by NIST and (D)ARPA

# Text REtrieval Conferences (TREC)

- TREC is an annual conference/competition devoted to information retrieval related tasks
- First TREC was held in 1992, since TREC-8 (1999) Q&A is one the tracks
- TREC is sponsored by NIST and (D)ARPA
- TREC provides human assessors to judge the answers

# TREC Q&A Track

# TREC Q&A Track

- Data Collection contained  $\approx 500,000$  ( $\approx 2\text{Gb}$ ) documents at TREC-8 and  $\approx 1,000,000$  ( $\approx 3\text{Gb}$ ) documents at TREC-9/10

# TREC Q&A Track

- Data Collection contained  $\approx 500,000$  ( $\approx 2\text{Gb}$ ) documents at TREC-8 and  $\approx 1,000,000$  ( $\approx 3\text{Gb}$ ) documents at TREC-9/10
- 200 (TREC-8), 693 (TREC-9), 500 (TREC-10) questions

# TREC Q&A Track

- Data Collection contained  $\approx 500,000$  ( $\approx 2\text{Gb}$ ) documents at TREC-8 and  $\approx 1,000,000$  ( $\approx 3\text{Gb}$ ) documents at TREC-9/10
- 200 (TREC-8), 693 (TREC-9), 500 (TREC-10) questions
- Answers were restricted to either 50 or 250 bytes

# TREC Q&A Track

- Data Collection contained  $\approx 500,000$  ( $\approx 2\text{Gb}$ ) documents at TREC-8 and  $\approx 1,000,000$  ( $\approx 3\text{Gb}$ ) documents at TREC-9/10
- 200 (TREC-8), 693 (TREC-9), 500 (TREC-10) questions
- Answers were restricted to either 50 or 250 bytes
- 36 groups participated at TREC-10 Q&A track (companies, universities, and governmental groups)

# Example Questions

# Example Questions

How many hexagons are on a soccer ball?

# Example Questions

How many hexagons are on a soccer ball?

Who played the teacher in Dead Poet's society?

# Example Questions

How many hexagons are on a soccer ball?

Who played the teacher in Dead Poet's society?

Where is the Kalahari desert?

# Example Questions

How many hexagons are on a soccer ball?

Who played the teacher in Dead Poet's society?

Where is the Kalahari desert?

When was Beethoven born?

# Example Questions

How many hexagons are on a soccer ball?

Who played the teacher in Dead Poet's society?

Where is the Kalahari desert?

When was Beethoven born?

Who is Desmond Tutu?

# Example Questions

How many hexagons are on a soccer ball?

Who played the teacher in Dead Poet's society?

Where is the Kalahari desert?

When was Beethoven born?

Who is Desmond Tutu?

Why did David Koresh ask the FBI for a word processor?

# Out of the Question

# Out of the Question

Which stocks should I buy?

# Out of the Question

Which stocks should I buy?

How do I partition my hard disk?

# Out of the Question

Which stocks should I buy?

How do I partition my hard disk?

Will Turkey join the EU?

# Out of the Question

Which stocks should I buy?

How do I partition my hard disk?

Will Turkey join the EU?

Does  $P = NP$ ?

# Out of the Question

Which stocks should I buy?

How do I partition my hard disk?

Will Turkey join the EU?

Does  $P = NP$ ?

Do you put a big stress on human resources management in Great Britain?

# Evaluation within the TREC Q&A Track

# Evaluation within the TREC Q&A Track

- For each question, a system is allowed to return 5 answers (ranked)

# Evaluation within the TREC Q&A Track

- For each question, a system is allowed to return 5 answers (ranked)
- Each answer has to be accompanied by the document it was taken from

# Evaluation within the TREC Q&A Track

- For each question, a system is allowed to return 5 answers (ranked)
- Each answer has to be accompanied by the document it was taken from
- If the correct answer is found at rank  $n$ , the score is  $1/n$ , otherwise it's 0

# Evaluation within the TREC Q&A Track

- For each question, a system is allowed to return 5 answers (ranked)
- Each answer has to be accompanied by the document it was taken from
- If the correct answer is found at rank  $n$ , the score is  $1/n$ , otherwise it's 0
- Best performing system at TREC-9 (50 bytes run) had an average score of  $\approx 0.6$

# Outline

- What's corpus-based Q&A ✓
- Q&A and TREC ✓
- Current approaches to Q&A
- The role of NLP in IR and Q&A
- The role of semantics in Q&A
- Our working plan
- Conclusions

# Outline

- What's corpus-based Q&A ✓
- Q&A and TREC ✓
- Current approaches to Q&A
- The role of NLP in IR and Q&A
- The role of semantics in Q&A
- Our working plan
- Conclusions

# Current approaches to Q&A

# Current approaches to Q&A

- Roughly speaking there are three kinds:

# Current approaches to Q&A

- Roughly speaking there are three kinds:
  1. Purely retrieval based (bag-of-words)

# Current approaches to Q&A

- Roughly speaking there are three kinds:
  1. Purely retrieval based (bag-of-words)  
Performance?

# Current approaches to Q&A

- Roughly speaking there are three kinds:
  1. Purely retrieval based (bag-of-words)  
Performance? rather badly ( $\approx 0.1 - 0.15$ )

# Current approaches to Q&A

- Roughly speaking there are three kinds:
  1. Purely retrieval based (bag-of-words)  
Performance? rather badly ( $\approx 0.1 - 0.15$ )
  2. Retrieval + named entity annotation

# Current approaches to Q&A

- Roughly speaking there are three kinds:
  1. Purely retrieval based (bag-of-words)  
Performance? rather badly ( $\approx 0.1 - 0.15$ )
  2. Retrieval + named entity annotation  
Performance?

# Current approaches to Q&A

- Roughly speaking there are three kinds:
  1. Purely retrieval based (bag-of-words)  
Performance? rather badly ( $\approx 0.1 - 0.15$ )
  2. Retrieval + named entity annotation  
Performance? okay ( $\approx 0.3$ )

# Current approaches to Q&A

- Roughly speaking there are three kinds:
  1. Purely retrieval based (bag-of-words)  
Performance? rather badly ( $\approx 0.1 - 0.15$ )
  2. Retrieval + named entity annotation  
Performance? okay ( $\approx 0.3$ )
  3. Retrieval + NE annotation + relation extraction

# Current approaches to Q&A

- Roughly speaking there are three kinds:
  1. Purely retrieval based (bag-of-words)  
Performance? rather badly ( $\approx 0.1 - 0.15$ )
  2. Retrieval + named entity annotation  
Performance? okay ( $\approx 0.3$ )
  3. Retrieval + NE annotation + relation extraction  
Performance?

# Current approaches to Q&A

- Roughly speaking there are three kinds:
  1. Purely retrieval based (bag-of-words)  
Performance? rather badly ( $\approx 0.1 - 0.15$ )
  2. Retrieval + named entity annotation  
Performance? okay ( $\approx 0.3$ )
  3. Retrieval + NE annotation + relation extraction  
Performance? best ( $\approx 0.6$ )

# What's in it for Q&A?

# What's in it for Q&A?

- There are only few studies that compare NLP based Q&A to other methods?

# What's in it for Q&A?

- There are only few studies that compare NLP based Q&A to other methods?
- Cardie et al. reported the following performance

# What's in it for Q&A?

- There are only few studies that compare NLP based Q&A to other methods?
- Cardie et al. reported the following performance

Technique	Mean Answer Rank

# What's in it for Q&A?

- There are only few studies that compare NLP based Q&A to other methods?
- Cardie et al. reported the following performance

Technique	Mean Answer Rank
Retrieval only	

# What's in it for Q&A?

- There are only few studies that compare NLP based Q&A to other methods?
- Cardie et al. reported the following performance

Technique	Mean Answer Rank
Retrieval only	3.07

# What's in it for Q&A?

- There are only few studies that compare NLP based Q&A to other methods?
- Cardie et al. reported the following performance

Technique	Mean Answer Rank
Retrieval only	3.07
+ Summarization	

# What's in it for Q&A?

- There are only few studies that compare NLP based Q&A to other methods?
- Cardie et al. reported the following performance

Technique	Mean Answer Rank
Retrieval only	3.07
+ Summarization	2.67 (+ 13.0%)

# What's in it for Q&A?

- There are only few studies that compare NLP based Q&A to other methods?
- Cardie et al. reported the following performance

Technique	Mean Answer Rank
Retrieval only	3.07
+ Summarization	2.67 (+ 13.0%)
+ NP parsing	

# What's in it for Q&A?

- There are only few studies that compare NLP based Q&A to other methods?
- Cardie et al. reported the following performance

Technique	Mean Answer Rank
Retrieval only	3.07
+ Summarization	2.67 (+ 13.0%)
+ NP parsing	2.66 (+ 13.4%)

# What's in it for Q&A?

- There are only few studies that compare NLP based Q&A to other methods?
- Cardie et al. reported the following performance

Technique	Mean Answer Rank
Retrieval only	3.07
+ Summarization	2.67 (+ 13.0%)
+ NP parsing	2.66 (+ 13.4%)
+ Semantic types	

# What's in it for Q&A?

- There are only few studies that compare NLP based Q&A to other methods?
- Cardie et al. reported the following performance

Technique	Mean Answer Rank
Retrieval only	3.07
+ Summarization	2.67 (+ 13.0%)
+ NP parsing	2.66 (+ 13.4%)
+ Semantic types	1.90 (+ 38.1%)

# Outline

- What's corpus-based Q&A ✓
- Q&A and TREC ✓
- Current approaches to Q&A ✓
- The role of NLP in IR and Q&A
- The role of semantics in Q&A
- Our working plan
- Conclusions

# Outline

- What's corpus-based Q&A ✓
- Q&A and TREC ✓
- Current approaches to Q&A ✓
- The role of NLP in IR and Q&A
- The role of semantics in Q&A
- Our working plan
- Conclusions

# The Role of NLP in Q&A

# The Role of NLP in Q&A

- Currently NLP is mainly used to

# The Role of NLP in Q&A

- Currently NLP is mainly used to
  - ▶ Do morphological normalization

# The Role of NLP in Q&A

- Currently NLP is mainly used to
  - ▶ Do morphological normalization
  - ▶ Detect phrases and entity types

# The Role of NLP in Q&A

- Currently NLP is mainly used to
  - ▶ Do morphological normalization
  - ▶ Detect phrases and entity types
  - ▶ Recognize relations between entities

# The Role of NLP in Q&A

- Currently NLP is mainly used to
  - ▶ Do morphological normalization
  - ▶ Detect phrases and entity types
  - ▶ Recognize relations between entities
- Requirements on NLP techniques

# The Role of NLP in Q&A

- Currently NLP is mainly used to
  - ▶ Do morphological normalization
  - ▶ Detect phrases and entity types
  - ▶ Recognize relations between entities
- Requirements on NLP techniques
  - ▶ Robust, robust, robust

# The Role of NLP in Q&A

- Currently NLP is mainly used to
  - ▶ Do morphological normalization
  - ▶ Detect phrases and entity types
  - ▶ Recognize relations between entities
- Requirements on NLP techniques
  - ▶ Robust, robust, robust
  - ▶ efficient

# The Role of NLP in Q&A

- Currently NLP is mainly used to
  - ▶ Do morphological normalization
  - ▶ Detect phrases and entity types
  - ▶ Recognize relations between entities
- Requirements on NLP techniques
  - ▶ Robust, robust, robust
  - ▶ efficient
  - ▶ accurate

# Morphological Normalization

# Morphological Normalization

- How to make sure that *company* and *companies* and *sell* and *sold* match?

# Morphological Normalization

- How to make sure that *company* and *companies* and *sell* and *sold* match?
- Matching morphological variants increases recall

# Morphological Normalization

- How to make sure that *company* and *companies* and *sell* and *sold* match?
- Matching morphological variants increases recall
- Morphological information such as tense and number has to be removed

# Morphological Normalization

- How to make sure that *company* and *companies* and *sell* and *sold* match?
- Matching morphological variants increases recall
- Morphological information such as tense and number has to be removed
- Two techniques to accomplish this:

# Morphological Normalization

- How to make sure that *company* and *companies* and *sell* and *sold* match?
- Matching morphological variants increases recall
- Morphological information such as tense and number has to be removed
- Two techniques to accomplish this:
  - ▶ Stemming

# Morphological Normalization

- How to make sure that *company* and *companies* and *sell* and *sold* match?
- Matching morphological variants increases recall
- Morphological information such as tense and number has to be removed
- Two techniques to accomplish this:
  - ▶ Stemming
  - ▶ Lemmatization

# Stemming

# Stemming

- A set of rules is used to remove suffixes

# Stemming

- A set of rules is used to remove suffixes
- E.g., plurals. If word ends in

# Stemming

- A set of rules is used to remove suffixes
- E.g., plurals. If word ends in
  - ▶ *ies*, but not *eies*, *aies*: *ies* → *y*
  - ▶ *es*, but not *aes*, *ees*, *oes*: *es* → *e*
  - ▶ *s*, but not *us*, *ss*: *s* →  $\emptyset$

# Stemming

- A set of rules is used to remove suffixes
- E.g., plurals. If word ends in
  - ▶ *ies*, but not *eies*, *aies*: *ies* → *y*
  - ▶ *es*, but not *aes*, *eas*, *oes*: *es* → *e*
  - ▶ *s*, but not *us*, *ss*: *s* →  $\emptyset$
- First applicable rule is the one used

# Lemmatization

# Lemmatization

- A lemmatizer reduces a word to its syntactic stem

# Lemmatization

- A lemmatizer reduces a word to its syntactic stem
  - ▶ Using morphological rules:

# Lemmatization

- A lemmatizer reduces a word to its syntactic stem
  - ▶ Using morphological rules:  
 $ies \rightarrow y,$

# Lemmatization

- A lemmatizer reduces a word to its syntactic stem
  - ▶ Using morphological rules:  
 $ies \rightarrow y, ed \rightarrow \emptyset,$

# Lemmatization

- A lemmatizer reduces a word to its syntactic stem
  - ▶ Using morphological rules:  
 $ies \rightarrow y, ed \rightarrow \emptyset, s \rightarrow \emptyset$

# Lemmatization

- A lemmatizer reduces a word to its syntactic stem
  - ▶ Using morphological rules:  
 $ies \rightarrow y, ed \rightarrow \emptyset, s \rightarrow \emptyset$
  - ▶ Unless a word is marked as irregular:

# Lemmatization

- A lemmatizer reduces a word to its syntactic stem
  - ▶ Using morphological rules:  
 $ies \rightarrow y, ed \rightarrow \emptyset, s \rightarrow \emptyset$
  - ▶ Unless a word is marked as irregular:  
 $sought \rightarrow seek,$

# Lemmatization

- A lemmatizer reduces a word to its syntactic stem
  - ▶ Using morphological rules:  
*ies* → *y*, *ed* →  $\emptyset$ , *s* →  $\emptyset$
  - ▶ Unless a word is marked as irregular:  
*sought* → *seek*, *sheep* → *sheep*,

# Lemmatization

- A lemmatizer reduces a word to its syntactic stem
  - ▶ Using morphological rules:  
*ies* → *y*, *ed* →  $\emptyset$ , *s* →  $\emptyset$
  - ▶ Unless a word is marked as irregular:  
*sought* → *seek*, *sheep* → *sheep*, *feet* → *foot*

# Lemmatization

- A lemmatizer reduces a word to its syntactic stem
  - ▶ Using morphological rules:  
 $ies \rightarrow y, ed \rightarrow \emptyset, s \rightarrow \emptyset$
  - ▶ Unless a word is marked as irregular:  
 $sought \rightarrow seek, sheep \rightarrow sheep, feet \rightarrow foot$
- Computationally more expensive since it requires a lexical look-up

# Named Entity Recognition

# Named Entity Recognition

- Recognize particular kinds of entities

# Named Entity Recognition

- Recognize particular kinds of entities
  - ▶ Persons

# Named Entity Recognition

- Recognize particular kinds of entities
  - ▶ Persons
  - ▶ Locations (cities, countries, states, regions, rivers, mountains, . . . )

# Named Entity Recognition

- Recognize particular kinds of entities
  - ▶ Persons
  - ▶ Locations (cities, countries, states, regions, rivers, mountains, . . . )
  - ▶ Companies

# Named Entity Recognition

- Recognize particular kinds of entities
  - ▶ Persons
  - ▶ Locations (cities, countries, states, regions, rivers, mountains, . . . )
  - ▶ Companies
  - ▶ Organizations

# Named Entity Recognition

- Recognize particular kinds of entities
  - ▶ Persons
  - ▶ Locations (cities, countries, states, regions, rivers, mountains, . . . )
  - ▶ Companies
  - ▶ Organizations
  - ▶ . . . .

# Named Entity Recognition

# Named Entity Recognition

- Techniques for recognizing named entities

# Named Entity Recognition

- Techniques for recognizing named entities
  - ▶ Gazetteers

# Named Entity Recognition

- Techniques for recognizing named entities
  - ▶ Gazetteers
  - ▶ Some form of feature learning

# Named Entity Recognition

- Techniques for recognizing named entities
  - ▶ Gazetteers
  - ▶ Some form of feature learning
- Gazetteers are problematic

# Named Entity Recognition

- Techniques for recognizing named entities
  - ▶ Gazetteers
  - ▶ Some form of feature learning
- Gazetteers are problematic
  - ▶ Inherently incomplete

# Named Entity Recognition

- Techniques for recognizing named entities
  - ▶ Gazetteers
  - ▶ Some form of feature learning
- Gazetteers are problematic
  - ▶ Inherently incomplete
  - ▶ Susceptible to causing false alarms

# Named Entity Recognition

- Techniques for recognizing named entities
  - ▶ Gazetteers
  - ▶ Some form of feature learning
- Gazetteers are problematic
  - ▶ Inherently incomplete
  - ▶ Susceptible to causing false alarms
- See the proceedings of the Message Understanding Conferences (MUC) for more sophisticated approaches

# Partial Parsing

# Partial Parsing

- Task: Identify phrases

# Partial Parsing

- Task: Identify phrases
- Mostly done by using finite-state techniques

# Partial Parsing

- Task: Identify phrases
- Mostly done by using finite-state techniques
  - ▶ Very efficient

# Partial Parsing

- Task: Identify phrases
- Mostly done by using finite-state techniques
  - ▶ Very efficient
  - ▶ For low level phrases (e.g., NP's, PP's, . . . ) ambiguity is not too big an issue

# Partial Parsing

- Task: Identify phrases
- Mostly done by using finite-state techniques
  - ▶ Very efficient
  - ▶ For low level phrases (e.g., NP's, PP's, . . . ) ambiguity is not too big an issue
- Prominent partial parsers are:

# Partial Parsing

- Task: Identify phrases
- Mostly done by using finite-state techniques
  - ▶ Very efficient
  - ▶ For low level phrases (e.g., NP's, PP's, . . . ) ambiguity is not too big an issue
- Prominent partial parsers are:
  - ▶ Cass (Abney)

# Partial Parsing

- Task: Identify phrases
- Mostly done by using finite-state techniques
  - ▶ Very efficient
  - ▶ For low level phrases (e.g., NP's, PP's, . . . ) ambiguity is not too big an issue
- Prominent partial parsers are:
  - ▶ Cass (Abney)
  - ▶ LT-Chunk (LTG, U Edinburgh)

# Partial Parsing

# Partial Parsing

*The celluloid torch has been passed to a new generation*

# Partial Parsing

*The celluloid torch has been passed to a new generation*

- NP's: [<sub>NP</sub>The celluloid torch] has been passed to [<sub>NP</sub> a new generation]

# Partial Parsing

*The celluloid torch has been passed to a new generation*

- NP's: [<sub>NP</sub>The celluloid torch] has been passed to [<sub>NP</sub> a new generation]
- PP's [<sub>NP</sub>The celluloid torch] has been passed [<sub>PP</sub> to [<sub>NP</sub> a new generation]]

# Partial Parsing

*The celluloid torch has been passed to a new generation*

- NP's: [<sub>NP</sub>The celluloid torch] has been passed to [<sub>NP</sub> a new generation]
- PP's [<sub>NP</sub>The celluloid torch] has been passed [<sub>PP</sub> to [<sub>NP</sub> a new generation]]
- VG's [<sub>NP</sub>The celluloid torch] [<sub>VG</sub> has been passed] [<sub>PP</sub>to [<sub>NP</sub> a new generation]]

# Verb Frames

# Verb Frames

- Argument-structures capture valuable information (*who did what to whom?*)

# Verb Frames

- Argument-structures capture valuable information (*who did what to whom?*)
- Fully parse and extract the argument structures

# Verb Frames

- Argument-structures capture valuable information (*who did what to whom?*)
- Fully parse and extract the argument structures
  - ▶ Computationally expensive

# Verb Frames

- Argument-structures capture valuable information (*who did what to whom?*)
- Fully parse and extract the argument structures
  - ▶ Computationally expensive
  - ▶ Robust?

# Verb Frames

- Argument-structures capture valuable information (*who did what to whom?*)
- Fully parse and extract the argument structures
  - ▶ Computationally expensive
  - ▶ Robust?
- Alternative: Given a VG and some NP's and/or PP's use the most probable matching frame one

# Verb Frames

- Argument-structures capture valuable information (*who did what to whom?*)
- Fully parse and extract the argument structures
  - ▶ Computationally expensive
  - ▶ Robust?
- Alternative: Given a VG and some NP's and/or PP's use the most probable matching frame one
  - ▶ Efficient,

# Verb Frames

- Argument-structures capture valuable information (*who did what to whom?*)
- Fully parse and extract the argument structures
  - ▶ Computationally expensive
  - ▶ Robust?
- Alternative: Given a VG and some NP's and/or PP's use the most probable matching frame one
  - ▶ Efficient, robust,

# Verb Frames

- Argument-structures capture valuable information (*who did what to whom?*)
- Fully parse and extract the argument structures
  - ▶ Computationally expensive
  - ▶ Robust?
- Alternative: Given a VG and some NP's and/or PP's use the most probable matching frame one
  - ▶ Efficient, robust, but more error prone

# Question Analysis

# Question Analysis

- Classify questions wrt answer type
  - ▶ How many hexagons are on a soccer ball?

# Question Analysis

- Classify questions wrt answer type
  - ▶ How many hexagons are on a soccer ball? (number)

# Question Analysis

- Classify questions wrt answer type
  - ▶ How many hexagons are on a soccer ball? (number)
  - ▶ Who played the teacher in Dead Poet's society?

# Question Analysis

- Classify questions wrt answer type
  - ▶ How many hexagons are on a soccer ball? (number)
  - ▶ Who played the teacher in Dead Poet's society? (person)

# Question Analysis

- Classify questions wrt answer type
  - ▶ How many hexagons are on a soccer ball? (number)
  - ▶ Who played the teacher in Dead Poet's society? (person)
  - ▶ Where is the Kalahari desert?

# Question Analysis

- Classify questions wrt answer type
  - ▶ How many hexagons are on a soccer ball? (number)
  - ▶ Who played the teacher in Dead Poet's society? (person)
  - ▶ Where is the Kalahari desert? (location)

# Question Analysis

- Classify questions wrt answer type
  - ▶ How many hexagons are on a soccer ball? (number)
  - ▶ Who played the teacher in Dead Poet's society? (person)
  - ▶ Where is the Kalahari desert? (location)
  - ▶ When was Beethoven born?

# Question Analysis

- Classify questions wrt answer type
  - ▶ How many hexagons are on a soccer ball? (number)
  - ▶ Who played the teacher in Dead Poet's society? (person)
  - ▶ Where is the Kalahari desert? (location)
  - ▶ When was Beethoven born? (date)

# Question Analysis

- Classify questions wrt answer type
  - ▶ How many hexagons are on a soccer ball? (number)
  - ▶ Who played the teacher in Dead Poet's society? (person)
  - ▶ Where is the Kalahari desert? (location)
  - ▶ When was Beethoven born? (date)
  - ▶ Who is Desmond Tutu?

# Question Analysis

- Classify questions wrt answer type
  - ▶ How many hexagons are on a soccer ball? (number)
  - ▶ Who played the teacher in Dead Poet's society? (person)
  - ▶ Where is the Kalahari desert? (location)
  - ▶ When was Beethoven born? (date)
  - ▶ Who is Desmond Tutu? (definition)

# Question Analysis

- Classify questions wrt answer type
  - ▶ How many hexagons are on a soccer ball? (number)
  - ▶ Who played the teacher in Dead Poet's society? (person)
  - ▶ Where is the Kalahari desert? (location)
  - ▶ When was Beethoven born? (date)
  - ▶ Who is Desmond Tutu? (definition)
  - ▶ Why did David Koresh ask the FBI for a word processor?

# Question Analysis

- Classify questions wrt answer type
  - ▶ How many hexagons are on a soccer ball? (number)
  - ▶ Who played the teacher in Dead Poet's society? (person)
  - ▶ Where is the Kalahari desert? (location)
  - ▶ When was Beethoven born? (date)
  - ▶ Who is Desmond Tutu? (definition)
  - ▶ Why did David Koresh ask the FBI for a word processor? (reason)

# Question Analysis

# Question Analysis

- Often, classifying questions is not non-ambiguous

# Question Analysis

- Often, classifying questions is not non-ambiguous
  - ▶ Who accused Microsoft of false advertising?

# Question Analysis

- Often, classifying questions is not non-ambiguous
  - ▶ Who accused Microsoft of false advertising?
  - ▶ Answer type can be: person, company, organization

# Question Analysis

- Often, classifying questions is not non-ambiguous
  - ▶ Who accused Microsoft of false advertising?
  - ▶ Answer type can be: person, company, organization
- Hierarchy of answer types

# Question Analysis

- Often, classifying questions is not non-ambiguous
  - ▶ Who accused Microsoft of false advertising?
  - ▶ Answer type can be: person, company, organization
- Hierarchy of answer types
  - ▶ Person, company, organization → named animate entity

# Question Analysis

- Often, classifying questions is not non-ambiguous
  - ▶ Who accused Microsoft of false advertising?
  - ▶ Answer type can be: person, company, organization
- Hierarchy of answer types
  - ▶ Person, company, organization → named animate entity
  - ▶ Time span and distance → length

# Question Analysis

# Question Analysis

- Identifying answer types requires a combination of

# Question Analysis

- Identifying answer types requires a combination of
  - ▶ (Partially) parsing the question

# Question Analysis

- Identifying answer types requires a combination of
  - ▶ (Partially) parsing the question
  - ▶ Pattern matching

# Question Analysis

- Identifying answer types requires a combination of
  - ▶ (Partially) parsing the question
  - ▶ Pattern matching
- Often fixed patterns occur in question

# Question Analysis

- Identifying answer types requires a combination of
  - ▶ (Partially) parsing the question
  - ▶ Pattern matching
- Often fixed patterns occur in question
  - ▶ *How long would it take to get from Earth to Mars?*

# Question Analysis

- Identifying answer types requires a combination of
  - ▶ (Partially) parsing the question
  - ▶ Pattern matching
- Often fixed patterns occur in question
  - ▶ *How long would it take to get from Earth to Mars?*
  - ▶ *What is the population of the Bahamas?*

# Question Analysis

- Identifying answer types requires a combination of
  - ▶ (Partially) parsing the question
  - ▶ Pattern matching
- Often fixed patterns occur in question
  - ▶ *How long would it take to get from Earth to Mars?*
  - ▶ *What is the population of the Bahamas?*
  - ▶ *What does Nicholas Cage do for a living?*

# Classification Patters

# Classification Patters

- Who is PERS-NAME? → PERS-DEFINITION

# Classification Patters

- Who is PERS-NAME? → PERS-DEFINITION
- Who is NP? → PERS-NAME

# Classification Patters

- Who is PERS-NAME? → PERS-DEFINITION
- Who is NP? → PERS-NAME
- When (did/was) . . . ? → DATE

# Classification Patters

- Who is PERS-NAME? → PERS-DEFINITION
- Who is NP? → PERS-NAME
- When (did/was) . . . ? → DATE
- How long ago . . . ? → TIME-PERIOD

# Classification Patters

- Who is PERS-NAME? → PERS-DEFINITION
- Who is NP? → PERS-NAME
- When (did/was) . . . ? → DATE
- How long ago . . . ? → TIME-PERIOD
- Who VP? → NAMED-ANIMATE

# Classification Patters

- Who is PERS-NAME? → PERS-DEFINITION
- Who is NP? → PERS-NAME
- When (did/was) . . . ? → DATE
- How long ago . . . ? → TIME-PERIOD
- Who VP? → NAMED-ANIMATE
- Where VP? → LOCATION

# Classification Patters

- Who is PERS-NAME? → PERS-DEFINITION
- Who is NP? → PERS-NAME
- When (did/was) . . . ? → DATE
- How long ago . . . ? → TIME-PERIOD
- Who VP? → NAMED-ANIMATE
- Where VP? → LOCATION
- What NP VP? → ???

# Answer Selection

# Answer Selection

- A question is represented as an argument structure QS with a question slot

# Answer Selection

- A question is represented as an argument structure  $QS$  with a question slot
- A phrase  $P$  from some argument structure  $AS$  is selected as answer if

# Answer Selection

- A question is represented as an argument structure  $QS$  with a question slot
- A phrase  $P$  from some argument structure  $AS$  is selected as answer if
  - ▶  $AS$  partially unifies with  $QS$

# Answer Selection

- A question is represented as an argument structure  $QS$  with a question slot
- A phrase  $P$  from some argument structure  $AS$  is selected as answer if
  - ▶  $AS$  partially unifies with  $QS$
  - ▶  $P$  is of the appropriate answer type, i.e.,  $P$  satisfies the constraints on the question slot

# Answer Selection

- A question is represented as an argument structure  $QS$  with a question slot
- A phrase  $P$  from some argument structure  $AS$  is selected as answer if
  - ▶  $AS$  partially unifies with  $QS$
  - ▶  $P$  is of the appropriate answer type, i.e.,  $P$  satisfies the constraints on the question slot
- Ranking answers requires weighted partial matching

# Outline

- What's corpus-based Q&A ✓
- Q&A and TREC ✓
- Current approaches to Q&A ✓
- The role of NLP in IR and Q&A ✓
- The role of semantics in Q&A
- Our working plan
- Conclusions

# Outline

- What's corpus-based Q&A ✓
- Q&A and TREC ✓
- Current approaches to Q&A ✓
- The role of NLP in IR and Q&A ✓
- The role of semantics in Q&A
- Our working plan
- Conclusions

# The Role of Semantics in Q&A

# The Role of Semantics in Q&A

- Extracting argument structure (semantics?)

# The Role of Semantics in Q&A

- Extracting argument structure (semantics?)
- Anaphora resolution

# The Role of Semantics in Q&A

- Extracting argument structure (semantics?)
- Anaphora resolution
- Synonym detection/Word-sense disambiguation

# Anaphora Resolution

# Anaphora Resolution

- Pronouns and definite NP's

# Anaphora Resolution

- Pronouns and definite NP's
  - ▶ Standard algorithms for resolving pronouns

# Anaphora Resolution

- Pronouns and definite NP's
  - ▶ Standard algorithms for resolving pronouns
  - ▶ Definite NP's are non-trivial (world-knowledge)

# Anaphora Resolution

- Pronouns and definite NP's
  - ▶ Standard algorithms for resolving pronouns
  - ▶ Definite NP's are non-trivial (world-knowledge)
  - ▶ Recognize definite NP's that do not co-refer

# Anaphora Resolution

- Pronouns and definite NP's
  - ▶ Standard algorithms for resolving pronouns
  - ▶ Definite NP's are non-trivial (world-knowledge)
  - ▶ Recognize definite NP's that do not co-refer  
*the world, the stock market, . . .*
- Relative pronouns

# Anaphora Resolution

- Pronouns and definite NP's
  - ▶ Standard algorithms for resolving pronouns
  - ▶ Definite NP's are non-trivial (world-knowledge)
  - ▶ Recognize definite NP's that do not co-refer  
*the world, the stock market, . . .*
- Relative pronouns
- Aliases

# Anaphora Resolution

- Pronouns and definite NP's
  - ▶ Standard algorithms for resolving pronouns
  - ▶ Definite NP's are non-trivial (world-knowledge)
  - ▶ Recognize definite NP's that do not co-refer  
*the world, the stock market, . . .*
- Relative pronouns
- Aliases
  - ▶ IBM vs. International Business Machines

# Anaphora Resolution

- Pronouns and definite NP's
  - ▶ Standard algorithms for resolving pronouns
  - ▶ Definite NP's are non-trivial (world-knowledge)
  - ▶ Recognize definite NP's that do not co-refer  
*the world, the stock market, . . .*
- Relative pronouns
- Aliases
  - ▶ IBM vs. International Business Machines
  - ▶ Hillary Rodham Clinton vs. Hillary Clinton

# Anaphora Resolution Example

# Anaphora Resolution Example

- *Who invented the paper clip?*

# Anaphora Resolution Example

- *Who invented the paper clip?*
- *The paper clip, weighing a desk-crushing 1,320 pounds, is a faithful copy of Norwegian Johan Vaaler's 1899 invention, . . . to honor the Norwegian who invented the office helper 90 years ago.*

# Synonyms/Hyponyms

# Synonyms/Hyponyms

- Matching the argument-structure of a question and an argument-structure from a document is often too strict

# Synonyms/Hyponyms

- Matching the argument-structure of a question and an argument-structure from a document is often too strict
  - ▶ Allow for partial matches

# Synonyms/Hyponyms

- Matching the argument-structure of a question and an argument-structure from a document is often too strict
  - ▶ Allow for partial matches
  - ▶ Allow synonyms/hyponyms to match

# Synonyms/Hyponyms

- Matching the argument-structure of a question and an argument-structure from a document is often too strict
  - ▶ Allow for partial matches
  - ▶ Allow synonyms/hyponyms to match
  - ▶ *Who invented the electric guitar?*

# Synonyms/Hyponyms

- Matching the argument-structure of a question and an argument-structure from a document is often too strict
  - ▶ Allow for partial matches
  - ▶ Allow synonyms/hyponyms to match
  - ▶ *Who invented the electric guitar?*  
*Adolph Rickenbacker was a Swiss immigrant who patented the first solid-body electric guitar*

# Synonyms/Hyponyms

- Matching the argument-structure of a question and an argument-structure from a document is often too strict
  - ▶ Allow for partial matches
  - ▶ Allow synonyms/hyponyms to match
  - ▶ *Who invented the electric guitar?*  
*Adolph Rickenbacker was a Swiss immigrant who patented the first solid-body electric guitar*
- Use WordNet for identifying pseudo-(syn/hyp)onyms

# Synonyms/Hyponyms

- Matching the argument-structure of a question and an argument-structure from a document is often too strict
  - ▶ Allow for partial matches
  - ▶ Allow synonyms/hyponyms to match
  - ▶ *Who invented the electric guitar?*  
*Adolph Rickenbacker was a Swiss immigrant who patented the first solid-body electric guitar*
- Use WordNet for identifying pseudo-(syn/hyp)onyms
- Optimally, the words are disambiguated

# Matching Examples

# Matching Examples

- 201 What was the name of the first Russian astronaut to do a spacewalk?

# Matching Examples

- 201 What was the name of the first Russian astronaut to do a spacewalk?
- 201 LA072490-0034 1 , Kubasov and Stafford , the commander of the Apollo mission on the flight . The broad - shouldered but paunchy Leonov , who in 1965 became the first man to walk in space , signed autographs .

# Matching Examples

- 201 What was the name of the first Russian astronaut to do a spacewalk?
- 201 LA072490-0034 1 , Kubasov and Stafford , the commander of the Apollo mission on the flight . The broad - shouldered but paunchy Leonov , who in 1965 became the first man to walk in space , signed autographs .
  - ▶ astronaut ISA man

# Matching Examples

- 201 What was the name of the first Russian astronaut to do a spacewalk?
- 201 LA072490-0034 1 , Kubasov and Stafford , the commander of the Apollo mission on the flight . The broad - shouldered but paunchy Leonov , who in 1965 became the first man to walk in space , signed autographs .
  - ▶ astronaut ISA man
  - ▶ spacewalk IS-RELATED-TO walk in space

# Matching Examples

- 201 What was the name of the first Russian astronaut to do a spacewalk?
- 201 LA072490-0034 1 , Kubasov and Stafford , the commander of the Apollo mission on the flight . The broad - shouldered but paunchy Leonov , who in 1965 became the first man to walk in space , signed autographs .
  - ▶ astronaut ISA man
  - ▶ spacewalk IS-RELATED-TO walk in space
  - ▶ Leonov ISA person

# Matching Examples

# Matching Examples

- 202 Where is Belize located?

# Matching Examples

- 202 Where is Belize located?
- 202 FBIS4-9714 1 Belize, with a population of 200,000, achieved its independence from the United Kingdom in 1986, and Guatemala has demanded the return of part of their common border since then.

# Matching Examples

- 202 Where is Belize located?
- 202 FBIS4-9714 1 Belize, with a population of 200,000, achieved its independence from the United Kingdom in 1986, and Guatemala has demanded the return of part of their common border since then.
  - ▶ border IS-RELATED-TO location

# Matching Examples

# Matching Examples

- 205 What is the population of the Bahamas?

# Matching Examples

- 205 What is the population of the Bahamas?
- 205 AP880314-0214 -1 s brother and ex-wife, of using a private airstrip to fly a 220-pound cocaine shipment to the Bahamas for delivery to Miami . Paul commands the powerful Dessalines Battalion that backs up police in this capital city with a population of 1 million .

# Matching Examples

- 205 What is the population of the Bahamas?
  - 205 AP880314-0214 -1 s brother and ex-wife, of using a private airstrip to fly a 220-pound cocaine shipment to the Bahamas for delivery to Miami . Paul commands the powerful Dessalines Battalion that backs up police in this capital city with a population of 1 million .
- ▶ **WRONG** candidate answer

# Matching Examples

# Matching Examples

- 205 What is the population of the Bahamas?

# Matching Examples

- 205 What is the population of the Bahamas?
- 205 FT931-15230 1 , as the Dollars 58m budgeted for it had been spent . Mr Ingraham 's charges of 'impropriety' are unlikely to excite the 245,000 people of the Bahamas . In his last decade in office , Sir Lynden was frequently forced to defend the government , and

# Matching Examples

- 205 What is the population of the Bahamas?
- 205 FT931-15230 1 , as the Dollars 58m budgeted for it had been spent . Mr Ingraham 's charges of 'impropriety' are unlikely to excite the 245,000 people of the Bahamas . In his last decade in office , Sir Lynden was frequently forced to defend the government , and
  - ▶ population is counted in people

# Matching Examples

# Matching Examples

- 217 What is the habitat of the chickadee?

# Matching Examples

- 217 What is the habitat of the chickadee?
- 217 LA012989-0070 1 A chain of lakes shimmer; chickadees carol in meadows at the foot of misty gabled peaks. The refulgent landscape is a balm and benediction

# Matching Examples

- 217 What is the habitat of the chickadee?
- 217 LA012989-0070 1 A chain of lakes shimmer; chickadees carol in meadows at the foot of misty gabled peaks. The refulgent landscape is a balm and benediction
  - ▶ landscape ISA location + habitat ISA location

# Matching Examples

# Matching Examples

- 221 Who killed Martin Luther King?

# Matching Examples

- 221 Who killed Martin Luther King?
- 221 AP881129-0181 -1 A memorial service honored Andrew Goodman, James Chaney and Michael Schwerner , civil rights workers killed in Mississippi in 1964, as well as others who were slain, such as Martin Luther King

# Matching Examples

- 221 Who killed Martin Luther King?
  - 221 AP881129-0181 -1 A memorial service honored Andrew Goodman, James Chaney and Michael Schwerner , civil rights workers killed in Mississippi in 1964, as well as others who were slain, such as Martin Luther King
- ▶ **WRONG ANSWER**

# Matching Examples

# Matching Examples

- 221 Who killed Martin Luther King?

# Matching Examples

- 221 Who killed Martin Luther King?
- 221 AP880401-0223 1 Earl Ray Says He 'd Trade Places with Hostages After nearly two decades in prison for the assassination of the Rev. Martin Luther King Jr. ,

# Matching Examples

- 221 Who killed Martin Luther King?
- 221 AP880401-0223 1 Earl Ray Says He 'd Trade Places with Hostages After nearly two decades in prison for the assassination of the Rev. Martin Luther King Jr. ,
- MLK Jr. and MLK refer to the same person???

# Matching Examples

- 221 Who killed Martin Luther King?
- 221 AP880401-0223 1 Earl Ray Says He 'd Trade Places with Hostages After nearly two decades in prison for the assassination of the Rev. Martin Luther King Jr. ,
- MLK Jr. and MLK refer to the same person???
- assassination NOMINALIZATION assassinate

# Matching Examples

- 221 Who killed Martin Luther King?
- 221 AP880401-0223 1 Earl Ray Says He 'd Trade Places with Hostages After nearly two decades in prison for the assassination of the Rev. Martin Luther King Jr. ,
- MLK Jr. and MLK refer to the same person???
- assassination NOMINALIZATION assassinate
- assassinate SYNONYM kill

# Matching Examples

# Matching Examples

- 318 Where did Bill Gates go to college?

# Matching Examples

- 318 Where did Bill Gates go to college?
- 318 FT923-14511 -1 By contrast, Bill Gates, the 36-year-old Harvard dropout and founder of Microsoft, may be the richest American, but he is only worth Dollars 6.4bn

# Matching Examples

- 318 Where did Bill Gates go to college?
- 318 FT923-14511 -1 By contrast, Bill Gates, the 36-year-old Harvard dropout and founder of Microsoft, may be the richest American, but he is only worth Dollars 6.4bn
  - ▶ dropout IS-RELATED-TO going to school

# Matching Examples

- 318 Where did Bill Gates go to college?
- 318 FT923-14511 -1 By contrast, Bill Gates, the 36-year-old Harvard dropout and founder of Microsoft, may be the richest American, but he is only worth Dollars 6.4bn
  - ▶ dropout IS-RELATED-TO going to school
  - ▶ school COORDINATE-TERM university

# Matching Examples

- 318 Where did Bill Gates go to college?
- 318 FT923-14511 -1 By contrast, Bill Gates, the 36-year-old Harvard dropout and founder of Microsoft, may be the richest American, but he is only worth Dollars 6.4bn
  - ▶ dropout IS-RELATED-TO going to school
  - ▶ school COORDINATE-TERM university
  - ▶ Harvard ISA university

# Matching Examples

- 318 Where did Bill Gates go to college?
- 318 FT923-14511 -1 By contrast, Bill Gates, the 36-year-old Harvard dropout and founder of Microsoft, may be the richest American, but he is only worth Dollars 6.4bn
  - ▶ dropout IS-RELATED-TO going to school
  - ▶ school COORDINATE-TERM university
  - ▶ Harvard ISA university
  - ▶ college COORDINATE-TERM university

# Outline

- What's corpus-based Q&A ✓
- Q&A and TREC ✓
- Current approaches to Q&A ✓
- The role of NLP in IR and Q&A ✓
- The role of semantics in Q&A ✓
- Our working plan
- Conclusions

# Outline

- What's corpus-based Q&A ✓
- Q&A and TREC ✓
- Current approaches to Q&A ✓
- The role of NLP in IR and Q&A ✓
- The role of semantics in Q&A ✓
- Our working plan
- Conclusions

# Our Working Plans

# Our Working Plans

- What got so far at ILLC:

# Our Working Plans

- What got so far at ILLC:
  - ▶ Most importantly, we got a name for our system:  
Tequesta

# Our Working Plans

- What got so far at ILLC:
  - ▶ Most importantly, we got a name for our system:  
Tequesta
  - ▶ IR system for pre-fetching relevant documents

# Our Working Plans

- What got so far at ILLC:
  - ▶ Most importantly, we got a name for our system:  
Tequesta
  - ▶ IR system for pre-fetching relevant documents
  - ▶ Part-of-speech tagger (TreeTagger, cudos to Helmut!)

# Our Working Plans

- What got so far at ILLC:
  - ▶ Most importantly, we got a name for our system:  
Tequesta
  - ▶ IR system for pre-fetching relevant documents
  - ▶ Part-of-speech tagger (TreeTagger, cudos to Helmut!)
  - ▶ Named entity recognizer

# Our Working Plans

- What got so far at ILLC:
  - ▶ Most importantly, we got a name for our system:  
Tequesta
  - ▶ IR system for pre-fetching relevant documents
  - ▶ Part-of-speech tagger (TreeTagger, cudos to Helmut!)
  - ▶ Named entity recognizer
  - ▶ A partial parser with verb frames (cudos to Sabine!)

# Our Working Plans

- What got so far at ILLC:
  - ▶ Most importantly, we got a name for our system:  
Tequesta
  - ▶ IR system for pre-fetching relevant documents
  - ▶ Part-of-speech tagger (TreeTagger, cudos to Helmut!)
  - ▶ Named entity recognizer
  - ▶ A partial parser with verb frames (cudos to Sabine!)
  - ▶ A question classifier

# Our Working Plans

# Our Working Plans

- What we still have to do:

# Our Working Plans

- What we still have to do:
  - ▶ Find appropriate weighting schemes for partial matching

# Our Working Plans

- What we still have to do:
  - ▶ Find appropriate weighting schemes for partial matching
  - ▶ Implement a resolver for anaphora

# Our Working Plans

- What we still have to do:
  - ▶ Find appropriate weighting schemes for partial matching
  - ▶ Implement a resolver for anaphora
  - ▶ Extract additional argument structures from nominalization (Nomlex)

# Our Working Plans

- What we still have to do:
  - ▶ Find appropriate weighting schemes for partial matching
  - ▶ Implement a resolver for anaphora
  - ▶ Extract additional argument structures from nominalization (Nomlex)
- Interesting question:

# Our Working Plans

- What we still have to do:
  - ▶ Find appropriate weighting schemes for partial matching
  - ▶ Implement a resolver for anaphora
  - ▶ Extract additional argument structures from nominalization (Nomlex)
- Interesting question:
  - ▶ To what extent does the quality of the NLP components affect the quality of the Q&A systems?

# Outline

- What's corpus-based Q&A ✓
- Q&A and TREC ✓
- Current approaches to Q&A ✓
- The role of NLP in IR and Q&A ✓
- The role of semantics in Q&A ✓
- Our working plan ✓
- Conclusions

# Outline

- What's corpus-based Q&A ✓
- Q&A and TREC ✓
- Current approaches to Q&A ✓
- The role of NLP in IR and Q&A ✓
- The role of semantics in Q&A ✓
- Our working plan ✓
- Conclusions

# Conclusions

# Conclusions

- NLP does make a difference in Q&A

# Conclusions

- NLP does make a difference in Q&A
  - ▶ Partial Parsing

# Conclusions

- NLP does make a difference in Q&A
  - ▶ Partial Parsing
  - ▶ Anaphora Resolution

# Conclusions

- NLP does make a difference in Q&A
  - ▶ Partial Parsing
  - ▶ Anaphora Resolution
  - ▶ Matching with synonyms

# Conclusions

- NLP does make a difference in Q&A
  - ▶ Partial Parsing
  - ▶ Anaphora Resolution
  - ▶ Matching with synonyms
- The exact impact is still a bit unclear

# Conclusions

- NLP does make a difference in Q&A
  - ▶ Partial Parsing
  - ▶ Anaphora Resolution
  - ▶ Matching with synonyms
- The exact impact is still a bit unclear
- Q&A is a nice framework for testing NLP tools in a real-world application