Question Answering

- Not the classic IR scenario
- Text Retrieval Conference TREC-8

  "People have questions and they need answers, not
  Automatic question answering will definitely be a
  advance in the state-of-art information retrieval te

- Unstructured document corpus
- Answers as portions of surface text
- Ranking of top 5 answers
- 50- and 250-byte versions

Approaches used in TREC-8

Information Retrieval

- Application of existing IR technology to the new problem
- Query processed for keywords
  $\rightarrow president, USA$
- Passages returned on statistical relevance measure tf/id
- Some performance with 250-byte passages
- Poor performance with 50-byte passages

Approaches used in TREC-8

Query Processing & Named Entity Extrac

- Query processed for keywords and required entity class

  Who $\rightarrow PERSON$
  When $\rightarrow TIME$
  How long $\rightarrow DURATION, LENGTH$

- Passages returned using IR techniques
- Entities within passages identified and assigned to semar

  Bill Clinton $\rightarrow PERSON$

- Various heuristics for choice of entity
- Most popular and best performing approach

Problems revealed by TREC-8

Q 1 *Who is the president of the USA?*

A 1.1 *Bill Clinton is the president of the USA.*

A 1.2 *Even if one disapproves of Bill Clinton, it is im member that although many people do not agree with h is one of the most powerful men on earth: he is the lea world and the president of the USA.*

A 1.3 *Hillary Clinton is the wife of the president of the*

Approaches used in TREC-8

Sentence Structure

- University of Maryland: dependency parser

  Dependency tree helps narrow entity choice
  Not enough to prevent selection of non-answers

- CL Research: semantic triples

  Semantic roles within sentence
  Restricted scope
  Full parser $\rightarrow$ not robust
  No coreference resolution

Aims of Project

- Use sentence structure to select answer entities and preve
- Requirements:

  Robust ($\rightarrow$ no full parse)
  Select portions of surface text as answer

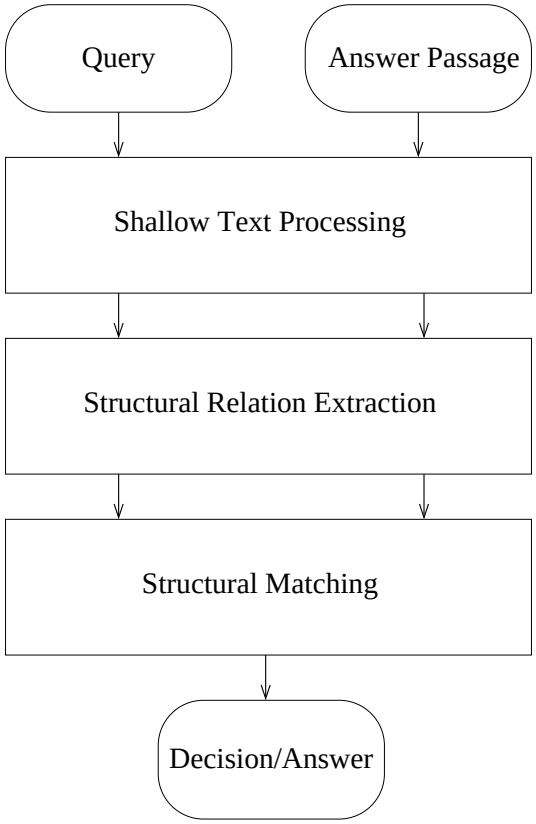
- Simplifications:

  Ignored tense, inference, negation

- Assumptions:

  IR query assumed (only passages containing keyword
  NEE assumed (entities added to lexicon)

System Overview



Shallow Text Processing

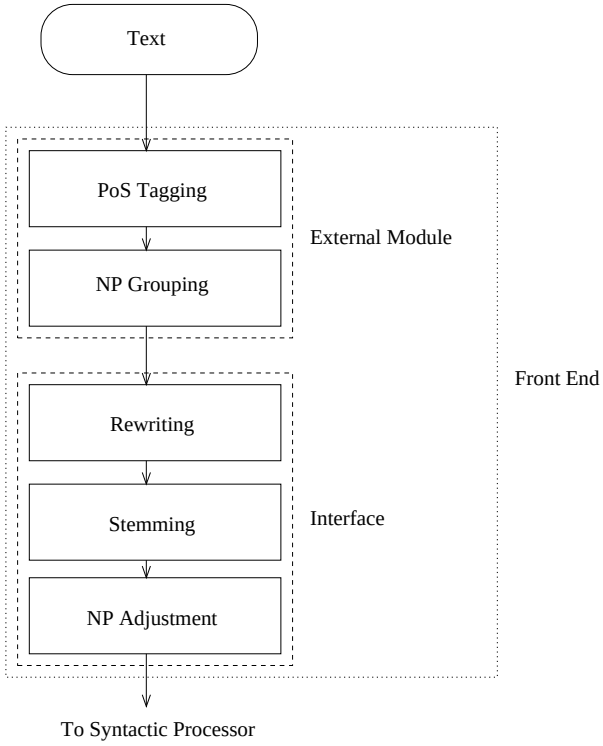


Figure 1: Front End

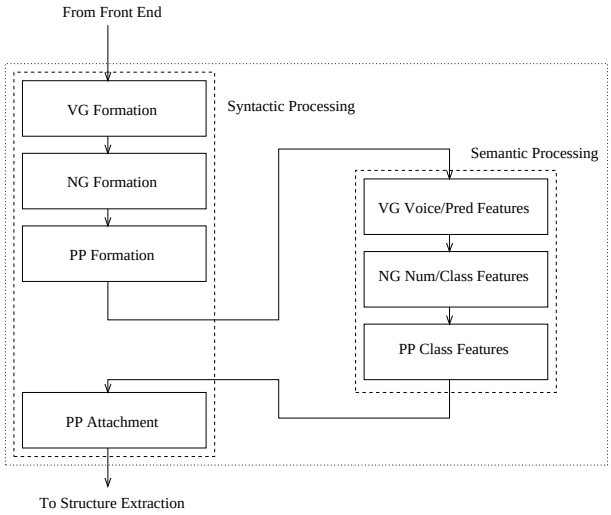


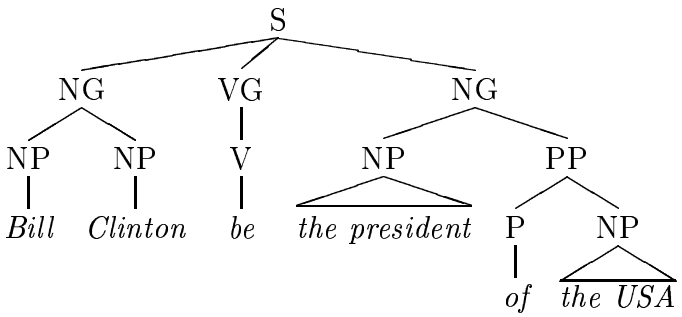
Figure 2: Main Functions

Syntactic Processing

Front End Tagging, stemming:

```
[np:[bill/nn],np:[clinton/np],be/vbz,np:[the
of/prep,np:[the/det,usa/np],'.'/'.']
```

VG/NG/PP Grouping Compound NPs, conjunctions, aux inversion



Semantic Processing

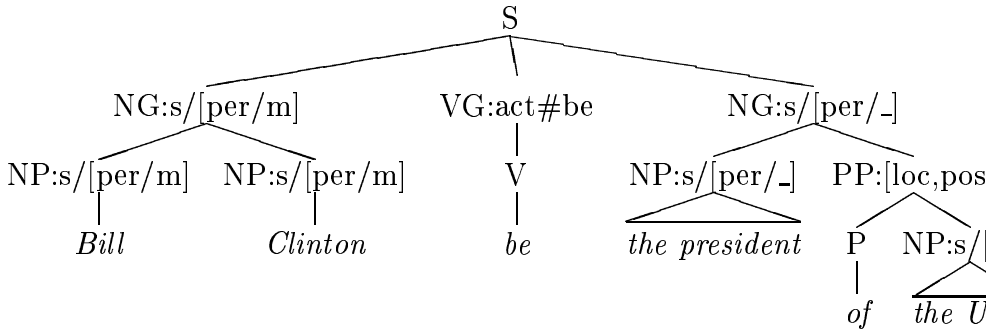
VG Info Predicate name, voice

NG/PP Info Semantic class, gender, number

Class	Description	Examples
`per/f`	Person (female)	queen, Mary
`per/_`	Person (either)	student, candidate
`obj`	Concrete object	banana, kitchen
`abs`	Abstract object	election, answer
`loc`	Location	Wales, Cambridge
`org`	Organisation	university, Microsoft

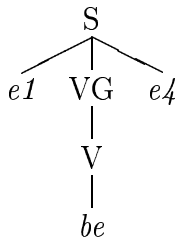
Class	Description	Examples	Re
`loc`	Location	in, outside	
`tim`	Time	in, before	
`man`	Manner	by	
`pos`	Possession	in, of	
`xxx`	Unknown	(anything)	

PP Attachment

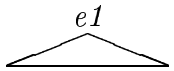


Coreference Resolution

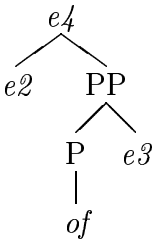
- Entity indexing - provide "pointers"



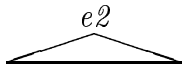
where



,



,



- Resolution of pronouns → NGs - by semantic class & nu
- Resolution of proper/definite NGs & number anaphora – ing more info
- Resolution of time & location anaphora
- Non-deterministic

Syntactic Simplification

- Complex sentences split into list of simple "equivalents"
- Conjunctions, punctuation

  "John likes Mary but Mary likes Bill"
  $\rightarrow$ *"John likes Mary"*, *"Mary likes Bill"*

- Subordinate, relative clauses

  "Mary, although John likes her, likes Bill"
  $\rightarrow$ *"Although John likes Mary"*, *"Mary likes Bill"*

  "Mary, who John likes, likes Bill"
  $\rightarrow$ *"John likes Mary"*, *"Mary likes Bill"*

- Queries & answers require different treatment later

Structure Extraction

- Predicate-argument structures formed from NG-VG-NG

```
s:[ like, john, mary ]
s:[ like, who, mary ]
```

- State structures formed if VG is existential, or from NG

```
s:[ [bill clinton], [[president],[of,[the
s:[ who, [[president],[of,[the usa]]] ]
```

Lexical Matching

- Query words matched with suitable entities by semantic type:

where	`loc`	PP
when	`tim`	PP, NG, NP
who, whom	`per`, `org`	NG, NP
what	`abs`, `obj`, `org`	NG, NP
which	`abs`, `obj`, `org`	NG, NP
whose	`pos`	PP
how	`man`	PP, AdvP
why	`rea`	PP

- VGs matched by predicate name
- NGs matched by content (complex NGs require many ru etc.)
- PPs matched by semantic class and NG matching (includi stacking)

Structural Matching

- Set of equivalence rules for tight matching of significantly
  tures
- Active-Passive

$$s : [VG_{pas}, Arg1, pp : [by, Arg2]] \Leftrightarrow s : [VG_{act}, Ar$$

- Existential Ordering

$$s : [Arg1, Arg2] \Leftrightarrow s : [Arg2, Arg1]$$

- Verb Nominalisation

$$s : [VG_{util}, Arg1, Arg2, \ldots] \Leftrightarrow s : [VG_{lemma}, Arg1',$$

$$\text{where} \quad Arg1' = Arg1 \setminus NP_{lemma}$$

- PP-Verbs (e.g. possession)

$$s : [VG_{pos}, Arg1, Arg2] \Leftrightarrow s : [Arg2, pp : [pos] : [o$$

Answer Passage Phenomena

Direct Match	*"Snowdon is in Wales"*
Singular/Plural	*"There are several racecourse in Newmarket"*
Unnecessary Modifiers	*"Snowdon, the highest mountain in the UK, is in Wales"*
Stem Matching	*"The biggest IT company is Microsoft"*
Inclusions	*"The Pacific islands like Hawaii produce tropical fruit like ba*
Expansions	*"In the Suffolk town of Newmarket there is a racecourse"*
Simple PPs	*"Livingstone is mayor of London"*
Possessive PPs	*"Ken Livingstone defeated all the other candidates and is Lo*
Verb Form Variations	*"Hawaii is producing bananas"*
VG Modifiers	*"Hawaii produced bananas in 1991"*
Verbs of Possession	*"Newmarket has a racecourse"*
Passives	*"One is made very fat by bananas"*
Existential Ordering	*"In Newmarket there is a racecourse"*
Existential Compounds	*"Snowdon, in Wales, is a serious mountain"*
Coreference	*"Pat is in the kitchen and Mike is there too"*
Conjunctions	*"Pat is in the lounge and Mike is in the garden"*
Relative Clauses (internal)	*"Hawaii specializes in the production of tropical fruit, which and pineapples"*
Relative Clauses (external)	*"Pat, who is reading, is in the lounge"*
Relative Clauses (coindexed)	*"Tropical fruit production, which includes banana production [. . .] of Hawaii"*
Subordinate Clauses	*"There is a library, with all the texts the student needs, in Cambridge"*
Verb Nominalisation	*"In Hawaii, which is not far from California, there is large s bananas"*

Evaluation

- Answers were assessed by a "notional user" with no detail ing of the system
- Portions of surface text marked manually and compare output
- Difficulty of consistency in marking → "narrow-scope" overlap allowed
- Conventional IR performance measures calculated:

$$Recall = n(correct \,\&\, identified)/n(correc$$

$$Precision = n(correct \,\&\, identified)/n(identi$$

- 86% recall, 97% precision on training data
- 50-80% recall, >90% precision in blind test (cf. TREC-8 5 answers)

Summary

- Structural relations used successfully, especially to reject
- Structural equivalences useful in the majority of cases
- Requires complex coding of rules, so large amounts of trai robust to unseen structural phenomena)

  "Who is the author of the book [. . .]?" ↔ *"[. . .] by*
- What do we do about inference & negation?

  "about 7 inches higher than 14,776 feet 2 inches" ↔ *inches"*