APPLICATION: DISCOURSE STRUCTURE AND QUESTION ANSWERING

SOFTWARE APPLICATION

OVERVIEW

- System for answering why-questions employing discourse relations in a pre-annotated document collection (the RST Treebank).
- Discourse structure can play an important role in complex question answering, but more forms of linguistic processing are needed for increasing recall.

WHY-QUESTIONS

- Why-questions have largely been ignored by researchers in the field of question answering (QA). One reason for this is that the frequency of why-questions posed to QA systems is lower than that of other types of question
- However as input for a QA system, they comprise about 5 percent of all wh-questions (Hovy et al., 2001; Jijkoun and De Rijke, 2005) and they do have relevance in QA applications (Maybury, 2003) such as who- and what-questions (Hovy et al., 2002)

- Techniques that have proven to be successful in QA for closed-class questions have been demonstrated to be not suitable for questions that expect an explanatory answer instead of a noun phrase (Kupiec, 1999).
- For why-QA on the other hand, more sophisticated techniques are needed, because most answers consist of some kind of reasoning that cannot be expressed in a noun phrase.

METHOD

- The paper follows Breck et al. (2000), which suggest that knowledge about discourse relations would have allowed their system for TREC-8 to answer why-questions.
- Research is limited to questions obtained from a number of subjects who were asked to read documents from the collection and formulate why-questions that another person would be able to answer given the text.

APPROACH FOR AUTOMATICALLY ANSWERING WHY-QUESTIONS

1. Question analysis and query creation
2. Retreival of candidate paragraphs or documents
3. Analysis and selection of text fragments
4. Answer generation.

ANSWER TYPES

- From (Verbene, 2006) is found that knowing the answer type of a question helps a QA system in selecting potential answers.
- Answer types for why-questions, based on Quirk et al. (1985): motivation, cause, circumstance and purpose. Of these, cause (52%) and motivation (37%) are by far the most frequent types in a set of why-questions pertaining to newspaper texts. With this syntax-based method, the correct answer type for 77.5% of these questions could be predicted(Verberne et al., 2006b).

RHETORICAL STRUCTURE THEORY

- **RST tree**
- The smallest units of discourse are called elementary discourse units (EDUs). A rhetorical relation typically holds between two EDUs:
- The nucleus is more essential for the writer's intention than the satellite. If two related EDUs are of equal importance, there is a multinuclear relation. Two or more related EDUs can be grouped together in a larger span, which in its turn can participate in another relation. By grouping and relating spans of text, a hierarchical structure of the text is created.

RST TREEBANK

- A treebank of manually annotated English texts with RST structures for testing purposes available.
- Created by (Carlson et al., 2003), contains a selection of 385 Wall Street Journal articles from the Penn Treebank that have been annotated with discourse structure in the framework of RST.
- The annotations are largely syntax-based, which fits the linguistic perspective of the research.
- Good levels of agreement have been measured between annotators of RST (Bosma, 2005)

HYPOTESIS

1. The question topic corresponds to a span of text in the source document and the answer corresponds to another span of text;
2. In the RST structure of the source text, an RST relation holds between the text span representing the question topic and the text span representing the answer.

Table 1. *Selected relation types*

Cause	Circumstance	Condition
Elaboration	Explanation-argumentative	Evidence
Interpretation	List	Problem-Solution
Purpose	Reason	Result
Sequence		

PROCEDURE

1. Identify the topic of the question.
2. In the RST tree of the source document, identify the span(s) of text that express(es) the same proposition as the question topic.
3. Is the found span the nucleus of a relation of one of the types listed in Table 1 (or, in case of cause relations, the satellite)? If it is, go to IV. If it is not, go to V.
4. Select the related satellite (or nucleus in case of a cause relation) of the found span as an answer.
5. Discard the current text span.

EXAPMLE

- Q: Why does Christopher Whittle think that Channel One will have no difficulties in reaching its target?
- Topic: Christopher Whittle thinks that Channel One will have no difficulties in reaching its target.
- Corresponding text span:
  "What we've done in eight weeks shows we won't have enormous difficulties getting to the place we want to be", said Mr. Whittle.

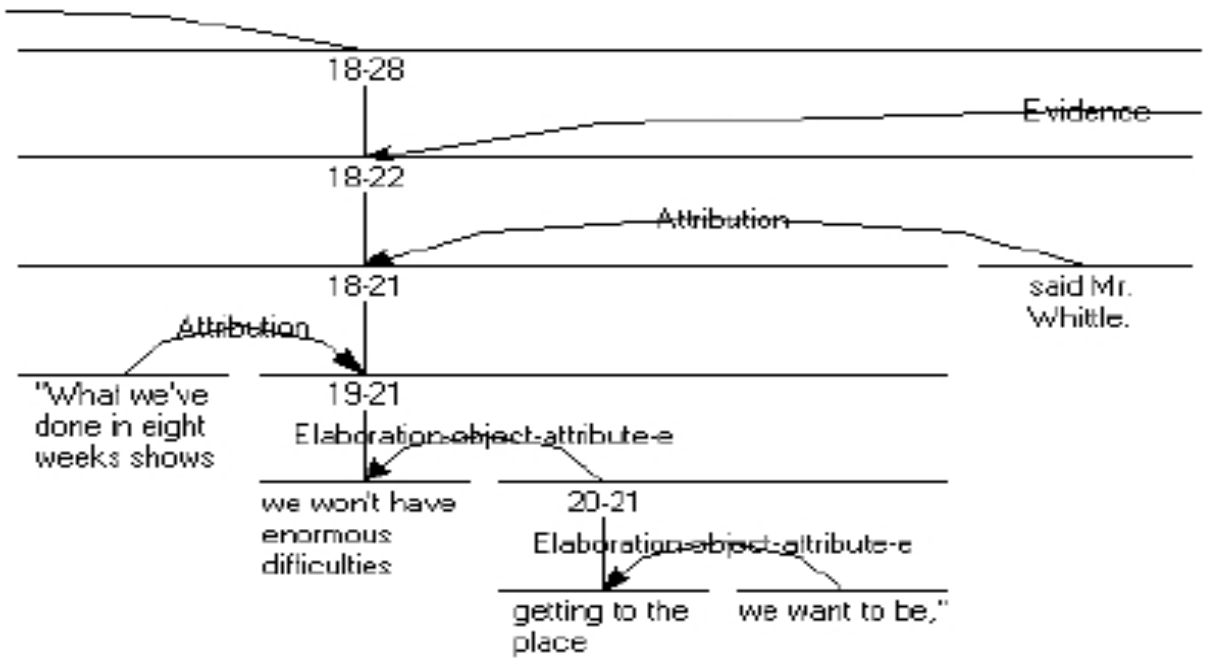


Figure 1. *RST sub-tree for the text span "What we've done in eight weeks shows we won't have enormous difficulties getting to the place we want to be, said Mr. Whittle."*

- A: He said his sales force is signing up schools at the rate of 25 a day. In California and New York, state officials have opposed Channel One. Mr. Whittle said private and parochial schools in both states will be canvassed to see if they are interested in getting the programs.

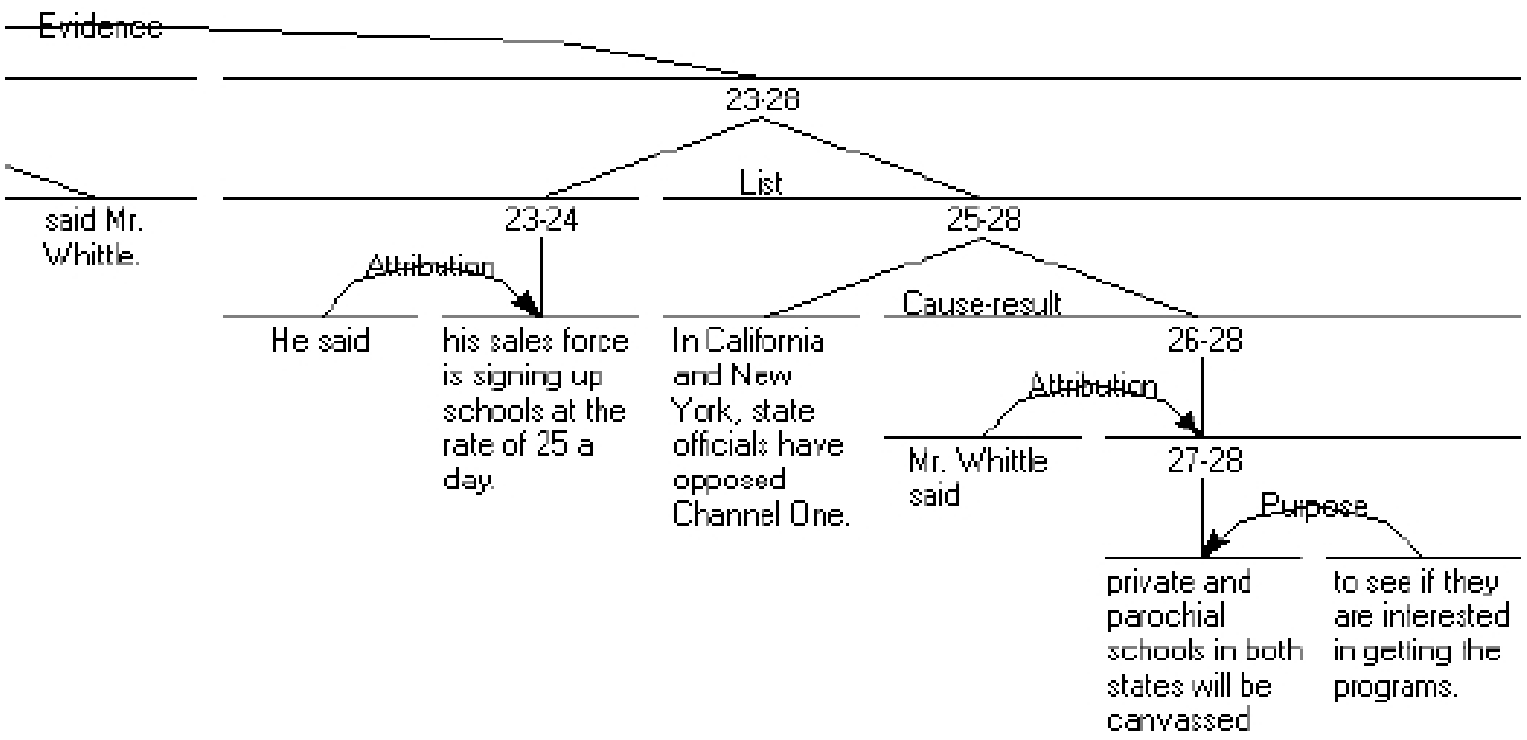


Figure 2. *RST sub-tree containing the satellite span "He said his sales force … to see if they are interested in getting the programs."*

IMPLEMENTATION

- The method was implemented in a Perl script. Most critical task is step 2: identifying the span(s) of text that express(es) the same proposition as the question topic.
- An Indexing script processes RST trees and searches for instances of relevant relations and saves them in an index file (for cause relations, nucleus and satellite are transposed).

```
( Nucleus (span 29 32) (rel2par span)
 ( Nucleus (leaf 29) (rel2par span) (text _!that interior regions of Asia
                                          would be among the first_!) )
 ( Satellite (span 30 32) (rel2par elaboration-object-attribute-e)
  ( Nucleus (leaf 30) (rel2par span) (text _!to heat up in a global warming_!) )
  ( Satellite (span 31 32) (rel2par consequence-n-e)
   ( Nucleus (leaf 31) (rel2par span) (text _!because they are far from
                                                    oceans,_!) )
   ( Satellite (leaf 32) (rel2par elaboration-additional) (text _!which
                                 moderate temperature changes.<P>_!) )
  )
 )
)
```

Figure 3. *Fragment of the original RST structure*

```
> consequence
   1. Nucleus (30): to heat up in a global warming
   2. Satellite (31 32): because they are far from oceans, which
                                 moderate temperature changes

> elaboration
   1. Nucleus (31): because they are far from oceans
   2. Satellite (32): which moderate temperature changes
```

Figure 4. *Fragment of the resulting index*

PROCEDURE

1. Read index file and normalize each nucleus in it. (removing all punctuation, lemmatization, applying a stop list, adding synonyms for each content word in the nucleus. Forms of normalization are combined into a number of configurations).
2. Read question and follow the same normalization procedure
3. Calculate likelihood
   (Relation prior is an indicator of the relevance of the relation type for why-questions and answer pairs)
4. Save all nuclei with a likelihood greater than the predefined threshold
5. Rank the nuclei according to their likelihood
6. For each of nuclei saved, print the corresponding answer satellite and the calculated likelihood.

Nucleus likelihood $P(N|Q) \sim P(Q|N) \cdot P(N)$

Question likelihood $P(Q|N) = \frac{\#\ question\ words\ in\ nucleus}{\#\ words\ in\ nucleus}$

Nucleus Prior $P(N) = \frac{1}{\#\ nuclei\ in\ document} \cdot P(R)$

Relation Prior $P(R) = \frac{\#\ instances\ of\ this\ relation\ type\ in\ question\ set}{\#\ occurrences\ of\ this\ relation\ type\ in\ treebank}$

RESULTS

MANUAL ANALYSIS

Table 2. *Outcome of manual analysis*

Question	# questions	% of questions
Questions analyzed	336	100
Questions for which we identified a text span corresponding to the topic	279	83.0
Questions for which the topic corresponds to the nucleus of a relation (or satellite in case of a cause relation)	207	61.6
Questions for which the satellite of this relation is a correct answer	195	58.0

In section 5.1, we will come back to the set of questions (42%) for which our procedure did not succeed.

EVALUATION

- The system was evaluated using outcome of manual analysis as a reference. Recall and MRR (1/rank of the reference answer, averaged over all questions) were measured
- Best performing is the configuration in which stop words are not removed, lemmatization is applied, no synonyms are added, and stop words and nonstop words are weighted 0.1/1.9
- Threshold added to reduce number of answers, based on the log probability is that our system calculates for each of the correct (reference) answers in our data collection. A probability that is slightly lower than the probabilities of these reference answers was chosen as threshold.

Table 3. *Main results for optimal configuration*

Recall (%)	53.3
Recall as proportion of questions for which the RST structure can lead to a correct answer (%)	91.8
Average number of answers per question	16.7
Mean reciprocal rank	0.662

Table 4. *Ranking of reference answer*

Answer rank	# questions	% of questions
Reference answer found	179	53.3
Reference answer ranked in 1st position	99	55.3
Reference answer ranked in 2nd to 10th position	60	33.5
Reference answer ranked in other position	20	11.2
Reference answer not found	157	46.7

QUESTIONS THAT COULD NOT BE ANSWERED USING THE METHOD:

- Questions whose topics are not or only implicitly supported by the source text. Half of them are supported by the text only implicitly, the other half is not supported at all by the text (World-knowledge needed).
- Questions for which topic and answer are supported by the source text but there is no RST relation between the span representing the question topic and the answer span (topic and answer refers to the same EDU, question topic and answer are embedded in different spans, often remote from each other).
- Questions for which the correct answer is not or only implicitly supported by the text.
- Questions for which the topic can be identified in the text and matched to the nucleus of a relevant RST relation, but the corresponding satellite is not suitable or incomplete as answer.

- Maximum recall that can be obtained from the use of RST relations as proposed in the present paper is 58.0%.
  Discarding the questions that require world knowledge, maximum recall is 73.9%.

COMPARISON WITH BASELINE

- A cue-based system was created to compare results with the RST Method.
- A system that follows this baseline method can obtain a maximum recall of 24.3% (4.5+2.2+17.6). This means that an RST-based method can improve recall by almost 140% compared to a simple cue-based method (58.0% compared to 24.3%).

SYSTEM SHORTCOMINGS

- There are 22 questions for which the manual analysis led to a correct answer, but the system did not retrieve this reference answer.
- For 17 of them, the nucleus was matched to the question manually, is not retrieved by the system because there is no (or too little, given the threshold) lexical overlap between the question and the nucleus that represent its topic.
- Some answers can be answered by using synonyms, increasing recall.
- The other 3 had inversed nucleus-satellite.

RELATIONS RELEVANCE

Table 5. *Addressed relation types*

Relation type	# referring questions	Relative frequency
Means	4	1.000
Purpose	28	0.857
Consequence	30	1.000
Evidence	7	0.750
Reason	19	0.750
Result	19	1.000
Explanation-argumentative	14	0.571
Cause	7	0.500
Condition	1	0.333
Interpretation	7	0.333
Circumstance	1	0.143
Elaboration	53	0.112
Sequence	1	0.091
List	4	0.016
Problem-Solution	0	0.000

RELATION TYPES CATEGORIES

- Relation types that are conceptually close of the general answer type reason ('core-why relations'): Purpose, Consequence, Evidence, Reason, Result, Explanation-argumentative and Cause. These relation types all have a relative frequency higher than 0.5 for why-questions.
- Relation types that are less applicable to why-questions ('non-why relations'): Means, Condition, Interpretation, Circumstance, Elaboration, Sequence, List and Problem-Solution.

- Considering the set of 207 questions for which the topic corresponds to the nucleus relation the recall is 77.5% (problematic cases excluded).
- This 207 questions were split into one set of questions whose answers can be found through a core-why relation (88.5 recall) and one sets of question whose answers that correspond to a non-why relation (60.3 recall)

EVALUATING ANSWER EXTRACTION FOR *WHY*-QA USING RST-ANNOTATED WIKIPEDIA TEXTS

INTRO

- Focus on the task of answer extraction for why-questions
- Performance of discourse-based answer extraction originated from real users' information needs.
- A corpus consisting of why-questions asked to the online QA system answers.com was created, and a set of manually selected Wikipedia fragments annotated manually with discourse structure was used.

QUESTION RETREIVAL

- Hovy et al. downloaded 17,000 questions from answers.com for their Webclopedia collection. 805 questions from the Webclopedia set are why-questions. The source of these questions guarantees that they originate from users' information needs.
- 400 of these why-questions were randomly selected for the data set. These questions were first studied independently from their answer documents.
- For 54% of the questions, the answer can be found in Wikipedia, of the other 46% some had false question propositions and other seemed to be too specific or too trivial.

NEW QUESTION CLASSES

- Motivation (10%), for example: Why did NBC reject the first \Star Trek" episode, \The Cage" in 1965?
- Physical Explanation (42%), for example: Why can't people sneeze with their eyes open?
- Non-physical explanation (30%), for example: Why is the color purple associated with royalty?
- Etymology (12%), for example: Why are chicken wings called Buffalo wings?
- Trivial/Nonsense (6%), for example: Why is the word “abbreviation” so long?

ANSWERS

- For analysis development a set of answer fragments for the 400 Webciclopedia why-questions were created manually from Wikipedia.
- In a large majority of cases (94%) the length of answer does not exceed a single paragraph.
- The answers were annotated by two experienced annotators. Inter-annotator agreement was measured for determining consistency. Moderate agreement on segmentation ($k = 0.54$) and low agreement for nuclearity ($k=0.13$) was obtained. Despite of the low results, annotations were still used.

DATA COLLECTION DIFFERENCE

- (a) the source of the questions (real user questions instead of elicited questions)
- (b) the source of the answer corpus (newly annotated encyclopedia fragments instead of pre-annotated newspaper texts)
- (c) the collection procedure (answers extracted for existing questions instead of questions formulated for existing answer documents).

QUESTIONS CONSIDERED

- For the experiment only questions which had answers in Wikipedia were considered
  60.6% of the questions were successfully answered, the remaining 39.4% suffers from one of the following problems:

1. The question topic is not represented by a text span in the answer fragment (18% of all questions);
2. The text span representing the question topic does not participate in an RST relation (2%);
3. The sibling of the span representing the question topic is not a satisfactory answer (21%).

RESULTS

Recall: 60.6%

Evaluating Answer Extraction for *Why*-QA using RST-annotated Wikipedia texts

Table 1.1: Distribution of relation types in corpora and question sets

	RST Treebank 37479 relations 372 *why*-questions		Wikipedia corpus 2333 relations 400 *why*-questions	
Relation type	% of relations	% of questions	% of relations	% of questions
Elaboration	18.0%	27.0%	22.4%	20.8%
Explanation	1.4%	7.1%	3.5%	20.0%
Circumstance	1.7%	0.5%	8.1%	16.0%
Background	0.5%	0.0%	4.3%	8.8%
Purpose	1.3%	14.3%	2.6%	7.2%
Consequence	1.0%	15.3%	0.8%	2.4%
Reason	0.6%	9.7%	0.9%	4.0%
Result	0.7%	9.7%	1.2%	2.4%

- Differences between categories are mainly originated by differences in annotation style and text source (news articles vs encyclopedia).
- Proportion of the question topics that participate in an elaboration relation for which this relation leads to a satisfactory answer: 49%.
- The predictive power for the question topics participating in an explanation-argumentative relation is much larger: 89%.
- For the question topics participating in a circumstance, background and purpose relation, these relations lead to a satisfactory answer in 77%, 85% and 100% of participating question topics respectively.
- Elaboration has low relevance.

SYSTEM RECALL

- Recall only of 25.9% on the Webclopedia/Wikipedia data set.
- Matching the question topic to spans in the source text using lexical overlap measures is hard since the questions are generated independently from the text.
- Approach should be combined with other methods to increase recall.

PARAGRAPH RETRIEVAL FOR WHY-QUESTION ANSWERING

- “How can we realize intelligent paragraph retrieval and paragraph ranking for why-QA, incorporating knowledge on discourse relations?”

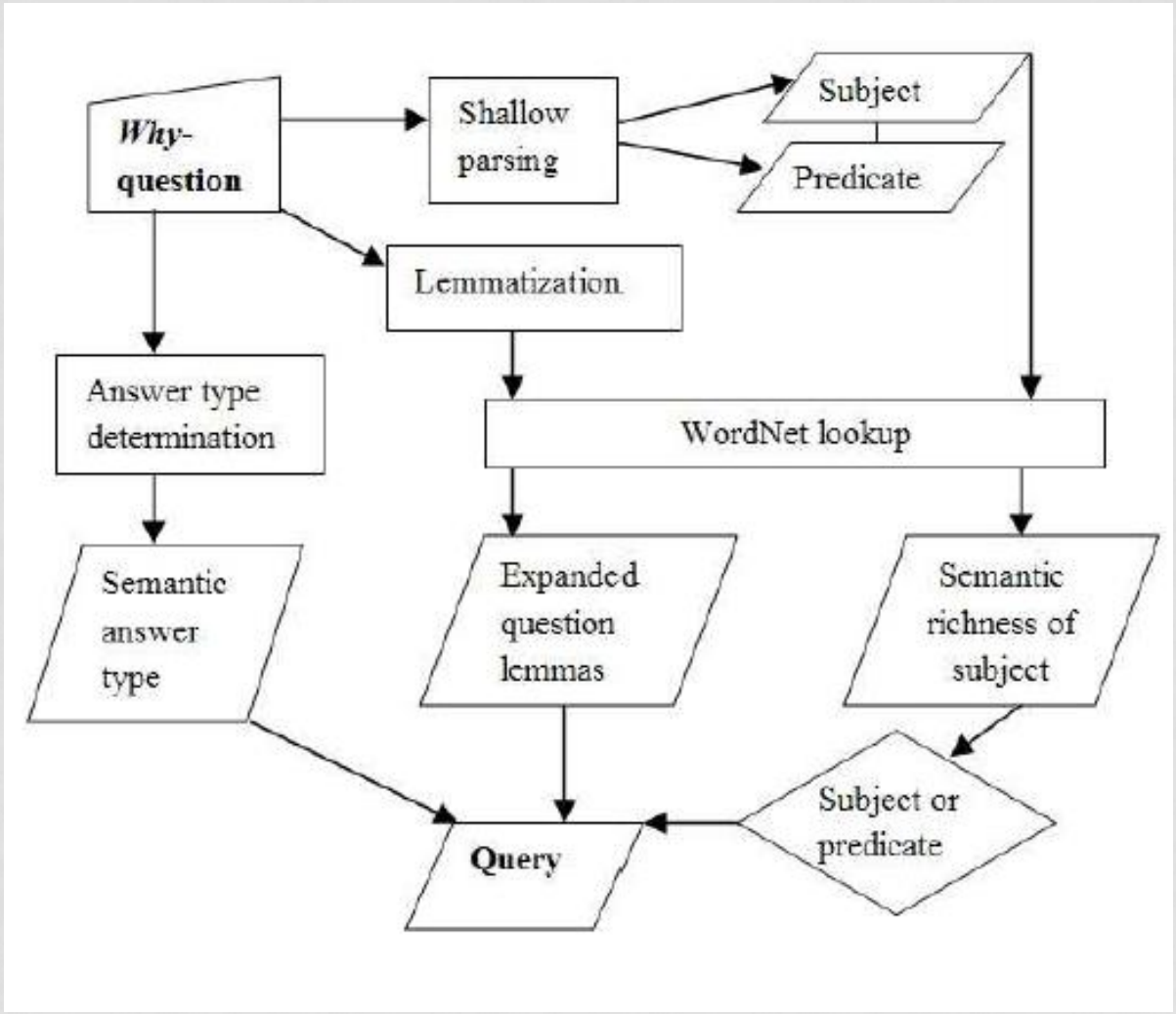
APPROACH

- (1) question analysis and query creation;
- (2) document retrieval;
- (3) paragraph retrieval and ranking.

METHOD

- Lexical expansion appeared to be very little help for retrieving discourse units, but it can probably play a more important role for questions formulated independently from the source text.
- The subject/predicate structure of the input question can be helpful in retrieving the answer document since the grammatical subject of a question often matcher the title of the answer document.
- If the subject is semantically rich, it is more likely it will lead to the answer documet.

- A method to divide subject and predicate of a why-question is needed.
- Shallow parsing together with fairly simple regular expression matching appears to be sufficient for this task. (Successful around 90% of the cases).


Why-question
Shallow parsing
Subject
Predicate
Lemmatization
Answer type determination
WordNet lookup
Semantic answer type
Expanded question lemmas
Semantic richness of subject
Subject or predicate
Query

- For ranking the retreived documents, the following variables were incorporated: lexical overlap between query and document text, lexical overlap between subject or predicate and document title, and a penalty for matching a synonym instead of a literal query term (threshold to be introduced).

SELECTION

- An intelligent approach to paragraph retreival and ranking is needed.
- An index for each of the parragraphs in a document must be created by the system including RST relations.

INFORMATION AVAILABLE

- (1) Information from the query
- (2) Title of the current document
- (3) Cues on the locations of potentially relevant RST relations in the text

- This variables (lexical, overlap, RST relations, document title) are needed to be combined into a probability model

ISSUES FOR DISCUSSION

- Retreival of short text fragments
- Intelligent paragraph retreival