Developing Offline Strategies for Answering Medical Questions

Erik Tjong Kim Sang, Gosse Bouma, Maarten de Rijke
University of Amsterdam, Rijksuniversiteit Groningen

July 10, 2005

IMIX: Interactive Multi-Modal Information Extraction

- Sponsor: Dutch research council NWO
- Time line: 2002–2007
- Goal: support developments for Dutch question answering
- Target: medical domain
- Keywords: speech recognition, dialogue management, multi-modal information presentation, information extraction
- Three groups work on question answering (since fall 2004)

Examples of target questions

- **What does RSI mean?**
  repetitive strain injury (answer format: named entity; question type: definition)

- **What causes RSI?**
  RSI is caused by repetitive tasks, awkward postures, forceful movements and insufficient resting times. (answer format: list; question type: cause)

- **What is a common misconception about RSI?**
  Many people believe that RSI is caused by working with a computer. Actually, there are many other activities that may cause RSI. RSI-related complaints have been frequently reported in professions like cleaners, welders, musicians and butchers. (answer format: paragraph; question type: other)

Frequency of question types and answer formats

Frequency	Question Type	Frequency	Answer Format
89	cause	157	list
73	prevention	148	yes/no
69	treatment	94	named entity
68	symptom	36	paragraph
55	definition		
19	diagnosis		
62	other		

Frequency of question types and answer formats in 435 RSI questions.

Example of layout-based IE from encyclopedic text

RSI

↑

(abbr. o. Repetitive Strain Injury), a series of disorders which can emerge when ...

Examples of professions with a higher chance on developing RSI are ...

Symptoms

↑

Two types of RSI exist. People that repeatedly perform the same...

People that perform lighter (office) work (e.g. journalists, computer operators) ...

Example of parsing-based IE from general text

- Cystic fibrosis is the most common inherited fatal disease
- A TIA is an early warning signal for a CVA
- Fatigueness can be identified from delayed responses
- Papillomes are caused by a virus
- Problems of the immune system can also cause rheumatism

Intrinsic evaluation of the IE techniques

	Precision		"Recall"	
Table type	**layout**	**parsing**	**layout**	**parsing**
definition	100%	18%	2870	7600
cause	99%	76%	333	6600
symptom	99%	78%	659	630
treatment	99%	n.a.	768	n.a.
prevention	99%	n.a.	56	n.a.
diagnosis	99%	n.a.	32	n.a.

Precision scores have been estimated by checking random samples from the output of the two information extraction techniques. The recall columns contain the number of derived patterns.

Extrinsic evaluation of the IE techniques

IE technique	correct	incomplete	wrong	missed
layout-based	11%	3%	70%	16%
parsing-based	4%	8%	88%	0%
layout, table only	5%	1%	6%	88%
human using corpus	18%	4%	0%	78%

Evaluation scores are based on the judgments by two assessors of answers generated for 50 randomly chosen questions.

Demo

A question answering demo based on the tables derived with layout-based information extraction is available on:

http://staff.science.uva.nl/∼erikt/factmine/qademo.cgi

Concluding remarks

- We have presented two approaches to information extraction for question answering in a restricted domain: one based on text layout and the other based on dependency parsing
- Both methods are reasonably precise but their recall is low
- Layout-based techniques require structured text of which we do not have a sufficient amount

Future work: continue with dependency parsing approach; incorporate online FAQ lists.

THE END