Question Answering in Biomedicine

Pierre Zweigenbaum

Mission de Recherche
en Sciences et Technologies de l’Information Médicale,
STIM/DPA/DSI, Assistance Publique - Paris Hospitals, Paris
& ERM 202, INSERM
pz@biomath.jussieu.fr
http://www.biomath.jussieu.fr/~pz/

Outline

☛ Task: Which Questions for Whom
    Task: Specificities of Medical Language Sources
    Resources: Health Information Sources
    Resources: Types of Questions
    Resources: Linguistic and Terminological Bases
    Conclusion

General orientation: Web-based Medical QA

Who

Health is one of the most frequent search topics on search engines

Who asks (bio)medical questions?

➢ Patients (the general public)
➢ Medical students
➢ Health care professionals
   family doctors (GPs)
➢ Biomedical researchers
➢ ...

Health Care Professionals

Half-life of medical knowledge is about 7 years
(2 years for biomedical knowledge)

➢ Continuous medical education
➢ Bibliographic databases (Medline, etc.)
➢ Direct Web search

Health Care Professionals: Example Questions

From (Ely et al., BMJ 2002)

☞ What is the dose of metformin?
☞ What is the proper treatment of gastro-oesophageal reflux disease (GERD)?
☞ What should I use for atopic dermatitis?
☞ How common is depression after infectious mononucleosis?
☞ What is the name of that rash that diabetics get on their legs?
☞ Is it ethical for me to take care of my own file clerk, who has back pain and wants a work excuse?

The General Public

➢ Before or after seeing a doctor
➢ For oneself or for relatives
➢ All about a given (suspected or diagnosed) disease,
   All about a given symptom,
   Indications and contraindications of treatment, etc.

Sometimes better informed on that topic than the doctor

Health Care Professionals: Time Issue

➢ Physicians spend less than 2 minutes on average seeking an answer to a question. (Ely et al., BMJ 1999)
➢ Thus, most clinical questions remain unanswered. (Alper et al., J Fam Pract 2001)
➢ Doctors are overwhelmed by the amount of information available, yet they often cannot answer their questions about specific clinical problems. (Ely et al., BMJ 2002)
➢ A niche for question-answering technology?
### Health Care Professionals: Main Obstacles

Main obstacles to answering doctor’s questions about patient care (Ely et al., BMJ 2002)

- **The excessive time** required to find information
- Difficulty modifying the original question
- Difficulty selecting an optimal strategy to search for information decide which resources will be most helpful; search in which order; which articles to read thoroughly; how thoroughly
- Failure of a seemingly appropriate resource to cover the topic
- Uncertainty about how to know when all the relevant evidence has been found so that the search can stop
- Inadequate synthesis of multiple bits of evidence into a clinically useful statement (e.g., conflicting evidence).

### Biomedical Researchers

- Search structured databases (FlyBase, SwissProt, etc.),
- Bibliographic databases (Medline, etc.),
- But also the Web faster update, more comprehensive

### Language Specialization: Terminological Variation

- **Variant denominations for a given disease or sign**
- Rich vocabulary: English terms, Neoclassical terms (Greek and Latin), eponyms, abbreviations...

<table>
<thead>
<tr>
<th>black lung disease</th>
<th>coal miners' lung</th>
</tr>
</thead>
<tbody>
<tr>
<td>coal workers' lung</td>
<td>miners' lung</td>
</tr>
<tr>
<td>miners' asthma</td>
<td>melanemia</td>
</tr>
<tr>
<td>anthracosis</td>
<td>pneumomelanosis</td>
</tr>
<tr>
<td>lung melanosis</td>
<td>coal miners' pneumoconiosis</td>
</tr>
<tr>
<td>coal workers' pneumoconiosis</td>
<td>coal pneumoconiosis</td>
</tr>
<tr>
<td>Collier's lung</td>
<td>Collier's anthracosis</td>
</tr>
</tbody>
</table>

### Language Specialization: Technicity

A spectrum of levels of technicity, which rely on diverse levels of medical knowledge:

- Specialist’s medical jargon (from specialist to specialist)
- General medical talk (from/to non-specialist)
- **Patient vocabulary** (“consumer vocabulary”) e.g., Medline Plus, CISMeF patient
  - Translate medical terms to patient terms (“synonym” approach)
  - Use medical terms and define them for patients (“glossary” approach)
  - Use different ontology for patients (Medline Plus)

### Trust: Reliability of Sources of Medical Information

- **Information used to make the most personal decisions, i.e., those concerning an individual’s health, should be of the highest quality.** (Hersh et al., JAMA 1998)
- Information about date and authorship is vital to assess the quality and currency of sources
- Even more stringent issue for QA since direct answers can be construed as implying a higher level of endorsement

- **The bulk of information on the World Wide Web [...] is of low applicability and poor quality for answering clinical questions.** (Hersh et al., JAMA 1998, based on 50 questions)
- Reliability is not taken into account by QA systems (Lin J., EACL, 12/4/2003, answer to question from the audience)
### Availability

**Selection bias of the Web**
- Computer science
- Trivia (geography and tourism, etc.)
- Health
- Health, reliable
- Health, French

### Low Availability

- Web redundancy may be much lower for technical, health questions
- Some questions have no answer on the Web
  (Jacquemart & Zweigenbaum, Stud Health Technol Inform 2003)
  found no answer through Google for 40% of a set of 100 student questions (oral pathology, French)
- Some (useful) questions have no known answer
  *Is smoking a risk factor for sinusitis?* (Ely et al., BMJ 2002)

### Need for Inference?

- Depends on distance between question and answering document
- Needs to be investigated

### Health Information Sources: Health Gateways

#### Quality-controlled health gateways
- **CISMeF** (Catalog and Index of French-language Health Resources, Rouen University Hospital, France) ([www.chu-rouen.fr](http://www.chu-rouen.fr))
  (Darmoni & Thirion, Meth Inf Med 2000)
- **CliniWeb** Oregon Health Sciences University, USA ([www.ohsu.edu/cliniweb](http://www.ohsu.edu/cliniweb))
  (Hersh et al., AMIA Annu Symp 1999)
- **DDRT** (Diseases, Disorders and Related Topics, [www.mic.ki.se/Diseases](http://www.mic.ki.se/Diseases)), Medical Library and Medical Information Center, Karolinska Institute, Stockholm, Sweden

#### Health Information Sources: Health Gateways

- **HON** (Health on the Net Foundation, Switzerland) ([www.hon.ch](http://www.hon.ch))
  (Boyer et al., Int J Med Inf 1997)
- **MedWebPlus** USA ([www.medwebplus.com](http://www.medwebplus.com))
- **OMNI** (Organizing Medical Networked Information UK) ([omni.ac.uk](http://omni.ac.uk))
  (Norman, Med Inf 1998)

#### Health Information Sources: CISMeF

- **CISMeF** (Darmoni et al., Meth Inf Med 2000)
  - Yahoo-type directory; 11,000+ “resources”
  - Primary intended audience: health care professionals but also general public, medical students
  - Selects resources according to quality criteria [www.medcertain.org](http://www.medcertain.org)
    - Addresses the issue of trust

---

**Task: Which Questions for Whom**
**Task: Specificities of Medical Language Sources**

- **Resources: Health Information Sources**
  - Resources: Types of Questions
  - Resources: Linguistic and Terminological Bases

**Conclusion**
Health Information Sources: Virtual Medical University

Primary intended audience: medical students

⇒ Online teaching material:
  Cap⇒ handbooks
  Cap⇒ exams and solutions
  Cap⇒ interactive pedagogical material
  Cap⇒ e.g., UMVF consortium of French medical universities (Le Beux et al., Inf Sant 2002)

Health Information Sources: Scientific Abstracts

Primary intended audience: health care professionals

⇒ NLM’s Medline (www.nlm.nih.gov)
⇒ Increasingly large number of free, full-text, online articles (Pubmed Central, BMJ, etc.)

Health Information Sources: Practice Guidelines

Primary intended audience: health care professionals

⇒ most up-to-date, “evidence-based” clinical knowledge
⇒ oncology, etc.
⇒ mostly textual knowledge bases (e.g., ANAES)

Health Information Sources: Consumer-Oriented Sites

Primary intended audience: general public

⇒ Governmental sites (e.g., NLM’s MedlinePlus, Québec’s Health Information site)
⇒ Associations of patients or families (e.g., Alzheimer’s disease)
⇒ Commercial sites (Doctissimo, etc.)

Health Information Sources: Knowledge Bases

Primary intended audience: health professionals; general public

⇒ Generally on CDROM; if Web, restricted access
⇒ Authoritative knowledge
⇒ Drug knowledge bases (e.g., Vidal, Physician’s Desk Reference)
⇒ Encyclopaedias
⇒ FAQs (e.g., www.dynamicmedical.com: GPs)

Task: Which Questions for Whom

Resources: Health Information Sources

⇒ Resources: Types of Questions

Resources: Linguistic and Terminological Bases

Conclusion

Health Information Sources: Knowledge Bases

⇒ Clinical questions by “family doctors”
⇒ Clinical questions by (practicing) medical students

Taxonomies of Medical Questions

A taxonomy of generic clinical questions (Ely et al., BMJ 2000)

⇒ 1396 questions collected from 152 family doctors (plus 14 general internists and 6 general paediatricians)
⇒ Goal-oriented classification: diagnosis, treatment, management, epidemiology, non-clinical
⇒ Four-level (max) hierarchy
⇒ A different perspective: contrasts with form-oriented classifications (Graesser, etc.), although sometimes congruent at inner levels (e.g., name finding)

A few generic question types account for a large number of questions
**Taxonomy of Generic Question Types (excerpts)**

**Diagnosis**

- cause/interpretation of clinical finding
  - What is the cause of \(X\)? (21.3%)
  - What is the differential diagnosis of symptom \(X\)?
  - What is the likelihood that symptom \(X\) is coming from condition \(Y\)?
  - At what level does physical finding \(X\) become clinically important?
  - How should I use test finding \(X\) in my decision?

**Treatment**

- Not only drug prescribing; efficacy/indications; treatment
  - How should I treat \(Y\) (given situation \(Z\))? (5.9%)
  - Should I use \(X\) for \(Y\)?
  - Is there any treatment for condition \(Y\)?
  - At what level of severity of condition \(Y\) is treatment indicated?

**Management**

- condition/finding
  - How should I manage \(Y\)? (4.8%)
  - What management options are there in situation \(Y\)?
  - How aggressive/conservative should I be in situation \(Y\)?

**Clinical Questions by Medical Students**

- 100 questions collected from medical students (oral pathology)
  - Quels signes caractérisent un lichen plan?
  - Doit-on traiter les atteintes buccales du lichen plan?
  - La mesure diagnostique une adénopathie?
  - The Reed-Sternberg cell evokes which disease?
Clinical Questions by Medical Students

- focus on whole triple (17%)
  Does diabetes contraindicate vasoconstrictors?
  Is it true that [A]–(complicates)–[B]?

- explanation on whole triple (5%)
  Pourquoi séparer un germe de dent de sagesse inclus?
  why [A]–(treats)–[B]

Total: 80% triple-based questions
The Unified Medical Language System

The UMLS project is a long-term NLM research and development effort designed to facilitate the retrieval and integration of information from multiple machine-readable biomedical information sources. (UMLS Documentation, 2002) umlsinfo.nlm.nih.gov

- Metathesaurus
- Semantic Network
- Specialist Lexicon

Distributed free of charge (sign convention with NLM); individual vocabulary sources may have usage restrictions

The UMLS Metathesaurus

The goal is to make it easier to develop systems that link information from patient record systems, bibliographic databases, factual databases, expert systems, etc. The UMLS Knowledge Sources can also facilitate the development of data creation and indexing applications. (UMLS Documentation, 2002)

- Houses the contents (terms, concepts and relations) of 100 medical terminologies (2003AA)
  - 800,000 concepts
  - 2,100,000 strings
- Identifies common concepts

What the Metathesaurus Is

- A repository of biomedical terms
- Addresses the issue of biomedical named entities
- A repository of families of “synonymous” terms
- Addresses the issue of terminological variation
- A huge network of interconnected concepts
- Addresses the issue of hierarchical navigation
- Not a terminology, but several overlayed terminologies
- Not an ontology: multiplicity of non-homogeneous structures

Biomedical Term Indexing with the UMLS

(Aronson et al., AMIA 2000, AMIA 2001)

Identify UMLS concepts in a document:

- ako Named Entity Recognition
- MetaMap program
- Cumulates the lexical variants of the Specialist Lexicon with the terminological variants of the Metathesaurus
- Used in the NLM Indexing Initiative

Can be obtained free of charge; no usage restrictions

Caveat: UMLS Non-English Content (terms)

e.g., French content?

- Metathesaurus: 25,000 concepts (2002)
  - i.e., 3% of Metathesaurus!
    - MeSH (29,173 strings)
    - WHO adverse reaction terminology (3,717)
    - digestive endoscopy (1,833)
    - primary care (723)

Currently, in French, UMLS terms differ from MeSH terms. Similar situation for Dutch, Finnish, Italian, Portuguese, Russian, and Spanish; German slightly better off (ICD-10).

- project VUMeF: add more French terms
Caveat: UMLS Non-English Content (lexicon)

Non-English medical lexicons?

- DSL: German Specialist Lexicon (inflection, derivation, compounding) (Weske-Heck et al., AMIA 2002)
- project UMLF: French Specialist Lexicon in preparation (inflection, derivation) (Zweigenbaum et al., MIE 2003)

In Need of an Ontology?

What is an ontology?

- The UMLS is not an ontology (does not support formal reasoning)
- GALEN ontology for medicine
- Gene Ontology for genomics

Synthesis: Challenges and Resources

- Challenges for question-answering systems
- Resources to address these challenges:
  - quality-controlled health information gateways offer a thorough indexing of trustworthy biomedical sources
  - taxonomies of question types rank and categorize interesting questions, taking into account their frequency of occurrence
  - biomedical lexicons, terminologies and ontologies are there to help manage domain-specific terms and concepts

The GALEN Ontology

A General Architecture for Language and Nomenclatures (Rector et al., MEDINFO 1992)

- Knowledge representation formalism (GRAIL, description logic)
- Define new concepts through composition
- Automatic classification
- Words or terms associated “externally” to concepts
- Sound, formal concept definitions; more complex than simple thesaurus
- Generative, potentially open; current coverage lower than UMLS

Concept definitions downloadable from www.opengalen.org

The Gene Ontology

- Gene product, molecular function, biological process, cellular component
- Terms, synonyms
- Multiple hierarchies, explicit hierarchical relations:
  - is-a
  - part-of
- Controlled vocabulary

Downloadable from www.geneontology.org

Perspective: Evaluation of Biomedical QA

- EQueR evaluation of French QA systems (Grau, 2003) in the context of the Technolangue EVALDA initiative for evaluating NLP systems (ELRA)
- (French) Medical QA track: participation welcome!

Acknowledgements

Many thanks to

- Aurel Craciun
- Tiphaine Dalmas
- Stéfan Darmoni
- Pierre Jacquemart