Towards Retrieving Relevant Information for Answering Clinical Comparison Questions

Annette Leonhard
School of Informatics
University of Edinburgh
EH8 9AB, Edinburgh, Scotland
annette.leonhard@ed.ac.uk

Abstract

This paper introduces the task of automatically answering clinical comparison questions using MEDLINE® abstracts. In the beginning, clinical comparison questions and the main challenges in recognising and extracting their components are described. Then, different strategies for retrieving MEDLINE® abstracts are shown. Finally, the results of an initial experiment judging the relevance of MEDLINE® abstracts retrieved by searching for the components of twelve comparison questions will be shown and discussed.

1 Introduction

Clinicians wishing to practice evidence-based medicine need to keep up with a vast amount of ever changing research to be able to use the current best evidence in individual patient care (Sackett et al., 1996). This can be difficult for time-pressed clinicians, although methods such as systematic reviews, evidence summaries and clinical guidelines can help to translate research into practice.

In a survey commissioned by Doctors.net.uk, 97% of doctors and nurses said that they would find a Question Answering (QA) Service useful, where they can ask questions in their own words (Bryant and Ringrose 2005). Studies have also shown that clinicians often want answers to particular questions, rather than getting information on broad topics (Chambliss & Conley, 1996; Ely et al., 1999, 2005).

A type of question that clinicians commonly want answered are comparison questions. In a corpus of clinical questions collected from the National Library of Health (NLH) Question Answering Service (http://www.clinicalanswers.nhs.uk), approximately 16% of the 4580 questions concern comparisons of different drugs, different treatment methods or different interventions as in (1).

> (1) Have any studies directly compared the effects of Pioglitazone and Rosiglitazone on the liver?

Despite the frequency of comparison questions in the clinical domain, there are no clinical QA methods specially designed to answer them. This paper introduces the task of answering clinical comparison questions, focusing initially on questions involving comparisons between drugs. Section 2 presents an overview of comparative structures and Section 3, relevant previous work on clinical question answering and the computational extraction of comparisons. Section 4 discusses strategies for retrieving MEDLINE® abstracts involving comparisons. Section 5 presents the results of an initial experiment judging the relevance of MEDLINE® abstracts, which are then discussed in Section 6.

2 Background

2.1 Indicators of Comparative Constructions

Proceedings of the Workshop on BioNLP, pages 153–161,
Boulder, Colorado, June 2009. ©2009 Association for Computational Linguistics

In order to identify questions about comparisons that should trigger special purpose search and extraction mechanisms, as well as identifying explicit comparisons made in text, one needs to recognize constructions commonly used to express comparisons in English (i.e. similarities and/or differences between two or more entities). In this paper, the term "entity" refers to drugs, treatment methods or interventions, and the initial focus of the work is on comparative questions in which two or more drugs or interventions are compared with respect to a particular criterion such as efficacy in treating a certain disease. This reflects their common occurrence in the NLH corpus.

Comparisons can appear in either a comparative form or a superlative form. The comparative form is used to compare two or more entities with respect to a certain attribute. The superlative form compares or contrasts one entity with a set of other entities and expresses the end of a spectrum. The following examples illustrate the difference:

Comparative form:
Is Ibuprofen **better** than Paracetamol for treating pain?
Superlative form:
Is Ibuprofen the **best** treatment for pain?

Friedman (1989) developed one of the first computational treatments of comparative structures. Comparisons are challenging because they correspond to a diverse range of syntactic forms such as coordinate or subordinate conjunctions, adverbial constructions or wh-relative-like clauses. Comparisons are cross-categorical and encompass adjectives, quantifiers, and adverbs. Adjectives and adverbs indicating comparisons occur in the following patterns:

Comparative adjectives and adverbs:

Regular adjectives and adverbs:
ADJ/ADV -er (e.g. safer) [[as/than][1] X] [for Y]
Irregular adjectives and adverbs:
e.g. worse/better [[as/than] X] [for Y]
Analytical adjectives and adverbs:
e.g. less/more *ADJ/ADV* [than X] [for Y]

Superlative adjectives and adverbs:

Regular adjectives and adverbs:
ADJ/ADV -est (eg. safest) X [for Y]
Irregular adjectives and adverbs:
e.g. worst/best X [for Y]
Analytical adjectives and adverbs:
e.g. least/most *ADJ/ADV* X [for Y]

Comparisons can also be expressed in other parts of speech. In the NLH corpus the following examples occur:

Verbs: compared to/with, differ from
Nouns: comparison, difference
Conjunctions: versus/vs, or and instead of

With respect to their semantics (and hence, with respect to other phrases or constructions they may appear with) comparatives can be *scalar* or *non-scalar* and express either *equality* or *inequality* between the compared entities. (Superlatives are absolute and the notion of scalability and equality does not apply to them).

Scalar adjectives and adverbs refer to attributes that can be measured in degrees, implying a scale along which entities can be arrayed. *Non-scalar* adjectives and adverbs refer to attributes that cannot be measured in degrees. *Equality* refers to constructs where two or more compared entities are equal in respect to a shared quality, whereas i*nequality* emphasises the difference between entities in respect to a certain quality.

Table 1 gives an example showing the four possibilities for drugs and interventions.

Scalability	Equality	Example
+	+	As efficient as x
-	+	Same intervention as x
+	-	Better treatment than x
-	-	Drug x differs from drug y

Table 1. Features of comparatives.

The difference between *scalar* and *non-scalar* comparisons plays an important role as far as automatic processing of comparative constructions with SemRep (Rindflesch and Fiszman, 2003; Rindflesch et al., 2005) is concerned. This will be discussed in Section 3.1.

[1] *As/ than* are optional. For example see "A or B: What is safer?"

Regular expressions based on the given patterns for adjectives and adverbs and on the other parts of speech shown above, as well as their respective part-of-speech tags, were used to extract a subset of comparison questions from a corpus collected from the National Library of Health Question Answering Service website at http://www.clinicalanswers.nhs.uk, as described in Section 2.3.

2.2 The NLH QA Service

The NLH Question Answering service (QAS) was a on-line service that clinicians in the UK could use to ask questions, that were then answered by a team of clinical librarians from Trip Database Ltd.[2], founded by Jon Brassey and Dr Chris Price. The questions and their answers were then retained at the website and indexed by major clinical topics (e.g. Cancer, Cardiovascular disease, Diabetes, etc.) so that clinicians could consult the QA archive to check whether information relevant to their own clinical question was already available.

While the NHS QAS service was discontinued in 2008, its archive of questions and answers was integrated into ATTRACT[3], the Welsh National Public Health Service run by Jon Brassey. The aim of both services has been to provide answers in a clinically relevant time frame using the best available evidence.

From the NLH QAS archive, a total of 4580 unique Q-A pairs of different degrees of complexity were collected for 34 medical fields representing questions asked and answered over a 36 month period. These were put into an XML format that separated the questions from the answers, while co-indexing them to indicate their association.

2.3 The Comparison Question Corpus

A sub-corpus specifically of comparison questions was created by POS-tagging the questions of the initial corpus with the Penn Treebank tagset, using the TnT tagger (Brants 1999). Regular expression were then used to search the tagged corpus for tagged lexical elements that indicated the constructions noted in Section 2.2.

Some questions were initially retrieved more than once because these questions contained more than one tag which was a comparison indicator. These duplicates were removed. There may be other comparative questions that might have been missed because of POS tagging errors. A small number of false positives were removed during manual post-processing. False positives were due to the fact that not all words tagged as superlatives are proper comparisons, but idiomatic expressions, such as "**best practise",** or proportional quantifiers (Huddleston and Pullum, 2002) such as "**Most** NSAIDs". (Scheible (2008) distinguishes eight different classes in which the superlative construction is used in English but only five of the eight classes involve true comparisons.) The result is a subset of 742 comparison questions out of the the total corpus of 4580 Q-A pairs.

Table 2. shows the number of occurrences for each item.

POS tag/Lexical item	Occurrences
JJR	195
RBR	124
JJS	207
RBS	68
versus, instead of	18
compared to/with, differ from	45
comparison, difference	85
Total	**742**

Table 2. Number of comparison indicators

3 Related Work

As the focus of this paper is biomedical text, the discussion here is limited to the work done in this context. Section 3.1 will present work on finding assertions involving comparisons in MEDLINE® abstracts and Section 3.2 will show work on answering clinical questions about comparisons.

3.1 Interpretation of Comparative Structures

(Fiszman et al., 2007) describes work on automatically interpreting comparative constructions in MEDLINE® abstracts. They use an extension of an

[2] http://www.tripdatabase.com/index.html

[3] http://www.attract.wales.nhs.uk/

existing semantic processor, SemRep (Rindflesch and Fiszman, 2003; Rindflesch et al., 2005), from the Unified Medical Language System resources to construct semantic predications for the extracted comparative expressions.

Fiszman et al. concentrate on extracting "structures in which two drugs are compared with respect to a shared attribute", such as a drug's efficacy in treating a certain condition, illustrated in the following in example:

(3) **Losartan** *was more effective than* **atenolol** in reducing cardiovascular morbidity and mortality in patients with hyptertension, diabetes, and LVH.
[Example (20) in (Fiszman et al. 2007)]

The drugs' relative merits in achieving their purpose is expressed by positions on a scale. Words like *than*, *as*, *with*, and *to* are cues for identifying compared terms, the comparison scale and the relative position of the compared entities on the scale.

Fiszman et al. focused on extracting the drug names, the scale and the position on the scale as illustrated in the SemRep representation from example (1):

(4) Losartan COMPARED_WITH Atenolol
Scale: Effectivness
Losartan HIGHER_THAN Atenolol
[Example (21) in (Fiszman et al. 2007)]

The overall F-score for the SemRep performance on the test set is 81% .

Fiszman et al. do not deal with questions, nor with identifying the basis of the comparison or the population in this paper, both of which are important for generating relevant answers for clinical questions. However, as Fiszman and Demner-Fushman have pointed out (personal communication), it is possible to identify the basis of the comparison and the population. Two drugs function as arguments to the TREATS predicate, which identifies the disease that is the basis for the comparison. SemRep can also identify the population using the predicate PROCESS_OF. For the question "Is treatment A better than treatment B for treating disease C in population D?", SemRep would produce the following representation for the basis of the comparison (C) and the population (D):

A TREATS **C**
B TREATS **C**
C PROCESS_OF **D**

There is an essential limitation to SemRep, however: Its comparative module only considers *scalar* comparative constructions, as presented in Section 2.1. *Non-scalar* comparisons, e.g. comparisons like "Is X the same intervention as Y?" or "How does drug X differ from drug Y?" cannot be extracted using SemRep. Also, the SemRep algorithm only recognises entities which occur on the left and the right side of the comparison cue and hence cannot recognize comparisons in which both compared entities are to the right side of the comparative cue as in "Which is better: X or Y?". This means that different methods are needed in order to process *non-scalar* comparisons and *scalar* comparisons that cannot be recognized because of their structure. In future work, rules will be defined for the different syntactic structures in which *non-scalar* comparisons and *scalar* comparison with both entities on the same side of a comparative cue can occur to serve as a basis for argument extraction during parsing.

There may also be problems with "Wh-" or "anything" questions (e.g. "What is better than X for treating Y?" or "Is there anything better than X for treating Y?"), because "Wh-words" or "anything" do not have a type that can be mapped. While Question Typing might solve such problems, the point is that questions involving comparisons raise somewhat different problems than assertions, which I will have to deal with in the work being carried out here.

3.2 Answering Clinical Questions

Demner-Fushman and Lin (2006) address superlative clinical questions of the type "What is the best treatment for X" by using a hybrid approach consisting of information retrieval and summarization.

Demner-Fushman and Lin's task breaks down into subtasks of identifying the drugs using UMLS concepts, clustering the abstracts for the drugs using UMLS semantic relationships and creating a short summary for each abstract by using the abstract title and outcome sentence. They focus primarily on synthesising correct answers from a set of search results consisting of MEDLINE® citations.

The system (*Cluster* condition) performs well compared to the baseline, which consists of the main interventions from the first three MEDLINE® abstracts retrieved by the manual PubMed queries. In a manual evaluation, only 20% of the drugs for the baseline were evaluated as beneficial, compared to 39% for the *Cluster* condition. 60% of the *PubMed* answers were judged as "good" in comparison to 83% for the *Cluster* condition.

The system orders the clusters by size, equating the most popular drug with the best drug. While this assumption is not always correct, the authors have observed that drugs that are studied more are more likely to be beneficial. In addition, while this approach might work for questions of the form "What is the best drug for X?" it cannot be used to answer other superlative questions such as Examples (5) or (6), because looking for the most studied drugs will not provide an answer to the question which drug has the fewest side effects or is safest to use.

(5) Which drug for treating X has the fewest side effects?

(6) Which drug is safest to use for treating X?

Despite this shortcoming, however, Demner Fushman and Lin's work of implementing an end-to-end QA system for superlatives provides a model for all future work in this area.

4 Strategies for Retrieving MEDLINE® Abstracts

As with (Fiszman et al., 2007) and (Demner-Fushman and Lin 2006), the current work starts with information retrieval. In particular, exploratory manual searches were first carried out via the OVID® portal to see if MEDLINE® abstracts are a useful resource for answering comparison questions such as "Is drug A better than drug B for treating X?"

With the assistance of a medical librarian from the University of Edinburgh's Information Services, different strategies to achieve the best possible retrieval of relevant abstracts were tried out. Two separate cases were considered: comparisons involving very popular, well-studied drugs and ones involving other drugs. First, strategies for the former will be described and illustrated with the following example question:

(7) Is paracetamol better than ibuprofen for reducing fever?

Titles and abstracts were searched for each compared entity (paracetamol and ibuprofen) and the basis of the comparison (fever). Then, the results were combined to return only abstracts containing both entities and the basis of the comparison. We found that search precision could be increased by limiting the search to *comparative study*, using OVID's publication type limit. That is, all abstracts that mention all three terms (i.e. the entities and the basis of the comparison) in the title or abstract involve relevant comparisons. The most common sources that were excluded by constraining the search to comparative studies are reviews, evaluation studies and case reports. These may contain relevant information but the initial focus was on the study type that was most likely to increase precision. (As the experiment reported in Section 5 and discussed in Section 6 shows, the restriction to comparative studies is insufficient to guarantee relevance.)

Constraining the search to comparative studies has somewhat different effects, depending on whether the drugs mentioned in the search are well-studied or not.

For popular, well-studied drugs, looking for the drug names often leads to hundreds of returned abstracts, most of which are not relevant. By including the basis of the comparison and limiting the study type to comparative studies, the number of returned abstracts for a set of 30 questions drops on average to 15% of the size of the original set of returned abstracts. For Example (7) a search for the combination of both drug names retrieved 593 abstracts. Including the basis of the comparison decreased the number to 139 abstracts. After constraining the results to comparative studies, the number of retrieved abstracts dropped to 24, which is a reduction of 83%.

For less-studied drugs, the difference in numbers of abstracts retrieved by including the basis of the comparison and limiting the search to the *comparative study* publication type is smaller compared

1. Is there any evidence to suggest that torasemide is better than furosemide as a diuretic?
2. Is lansoprazole better than omeprazole in treating dyspepsia?
3. Are there any studies comparing topical diclofenac gel with ibuprofen gel?
4. Effectiveness of Decapeptyl in treatment of prostate cancer in comparison to Zoladex?
5. Which is more effective ibuprofen or diclofenac for arthritis pain for pain relief?
6. Is calcium citrate better absorbed and a more effective treatment for osteoporosis than calcium carbonate?
7. Have any studies directly compared the effects of Pioglitazone and Rosiglitazone on the liver?
8. Is Famvir (famciclovir) better than acyclovir for Herpes zoster?
9. Is it true that men on captopril have a better quality of life than men on enalapril?
10. What is the first choice for Type 2 diabetes patients: sulphonylurea or metformin?
11. Is there any evidence as to which is more effective at preventing malaria: Malarone or Doxycyline?
12. In conjunctivitis which is better chloramphenicol or fucithalmic eye drops?

Figure 1. Questions used in the experiment.

to the numbers retrieved by only looking for the drug names, because fewer abstracts exist for these drugs, but the relevance of the returned abstracts improves as considerably as for the more studied drugs. (Recall was not analyzed during the explorations because for answering clinical questions the relevance of the retrieved abstracts is more important than retrieving all possible abstracts.)

There have also been cases where including the basis of the comparison leads to the return of no relevant abstracts. In this case, different strategies from the one discussed above will be necessary.

Often drugs are known under generic names or the basis of the comparison is related to symptoms which are not explicitly mentioned in the question but which are still relevant. In order to recognise that different terms are actually related to the same drug or disease and belong to the same hierarchy, advantage was taken of OVID's ability to map the entities to their corresponding MeSH (Medical Subject Headings) terms and to "explode" the MeSH terms to include all of the narrower, more specific subheadings during the search.

So far the focus has been on manual retrieval of abstracts. The described search strategy of combining search terms and restricting the results to the specific publication type could have been done using a search engine which implements Boolean operators and is capable of indexing XML documents However, the description of the search strategy and the presentation of the intermediate searches, which would have been performed internally by a search engine, was regarded important to illustrate the impact of adding the basis of the comparison and the use of a publication type limit on the number of retrieved abstracts.

5 Judging the Relevance of MEDLINE® Abstracts

A initial experiment was carried out to evaluate the relevance of the abstracts retrieved from MEDLINE® via Ovid® using the strategies described in the previous section.

The experimental subjects were eight 4th year medical students, who evaluated the abstracts retrieved for twelve clinical comparison questions in which two drugs were compared to each other with respect to a particular attribute. The questions differ in syntactic structure, but they all contain comparisons of two drugs. Figure 1 shows the list of questions.

The material presented to the medical students in the experiment was created as follows: The drug names and the basis of the comparison from the natural language questions were manually mapped to their corresponding MeSH terms and used to retrieve abstracts via OVID® using the final strategy described in Section 4.

For any question, the maximum number of abstracts given to the student judges was 15, comprising up-to-15 of the most recent abstracts. In total, each judge evaluated 103 abstracts. Each abstract was assigned by each judge into one of three categories, based on the criteria given after the category label:

1. Relevant: Both drugs from the question or their generic names are mentioned in the abstracts, the drugs are directly compared to each other and the disease or the attribute with respect to which they are being compared is also mentioned and the same as stated in the question or synonymous to it (e.g. heartburn and dyspepsia would both count as right because they are closely related).

2. Not Relevant: The drugs or their generic names are not mentioned in the abstract, the drugs are not compared and/or the disease or the attribute with respect to which they are being compared is wrong (as in different from what is stated in the question, e.g. effect on blood pressure instead of use as a painkiller).

3. Somewhat Relevant: The drugs or their generic names are mentioned but there are no single sentences indicating a comparison between them or the disease is not mentioned. If the wrong disease is mentioned, the abstract should be labeled "not relevant".

The judges were also asked to explain the reason for their choice of labels.

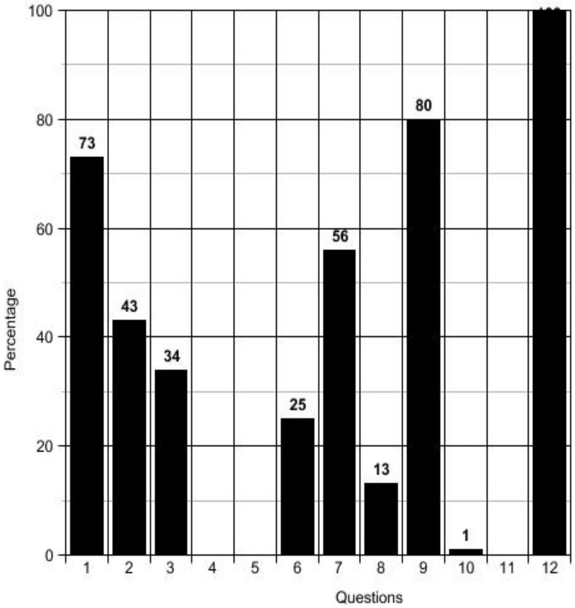
The inter-annotator agreement between the judges was computed using a variant kappa statistic for multiple annotators (Fleiss, 1971). The null hypothesis was rejected and it was ensured that the observed agreement is not accidental.

Overall inter-annotator agreement for all three categories measured by the kappa statistic was moderate at 0.58 for a total of 103 judgments. 47 judgments were in the "somewhat relevant" category. If annotator agreement is only assessed on the remaining 56 judgments from the two categories "relevant" and "not relevant", kappa is 0.97, which represents almost perfect agreement.

6 Results and Discussion

Graph 1 shows the percentage of abstracts that were judged relevant by the eight judges for each question. The numbers of retrieved abstracts for each question were: 15 abstracts for Question 1, 5, 8 and 10, 9 abstracts for question 7 and 11, 7 abstracts for Question 2, 5 abstracts for Question 9, 4 abstracts for Question 6 and 12, 3 abstracts for Question 3 and 2 abstracts for Question 4.

Question 1, 9 and 12 show a very high percentage of relevant abstracts (73%, 80% and 100% respectively), whereas no relevant abstracts were retrieved for questions 4, 5 and 11, and only one relevant abstract (out of 15) for question 10. An abstract was considered relevant when at least five of the eight judges considered it relevant.



Graph 1. Percentage of abstracts judged relevant by the majority of the judges for each of the twelve questions. The label on the top of each bar is the actual percentage.

Here the main sources for these disparate results are discussed, based on both the explanations given by the student judges and discussions with our medical librarian.

Approximately 30% (31 of 103) of the abstracts were labeled "not relevant" by the judges because they lacked any direct evidence of a comparison e.g. at least one sentence that explicitly compares the two drugs in question, even though the drugs are mentioned in the abstract and the study is a *comparative study* (as indicated in its MeSH indices). This is illustrated in Example (9), which shows the three sentences from one of the abstracts retrieved for Question 1 that explicitly mention the two drugs:

> (9) Piretanide and **furosemide** have a constant extrarenal elimination and thus accumulate in renal failure.[...] Elimination of **torasemide** is independent of its renal excretion. Thus in renal failure, **torasemide** is the only loop diuretic in which the plasma concentration is strictly dose dependent.

About 10% (10) of the abstracts were judged to be irrelevant because the drugs were compared as part of a treatment regime in combination with other drugs, as in Abstract 4 for Question 6 in which calcium citrate and calcium carbonate are compared co-administered with different preparations of sodium fluoride. In two cases (2% of the abstracts), doses of a given drug were compared against other dosages instead of the drugs themselves, e.g. 30 mg lansoprazole versus 20mg omeprazole.

A major factor for "not relevant" judgments was the time frame. This was relevant when retrieving abstracts about well-established drugs that have been in existence for a long time, such as ibuprofen or diclofenac. All but one of the 18 abstracts retrieved for the two questions about these two drugs were irrelevant, even though the two drugs were explicitly mentioned in the abstract. The problem is that they were grouped together as conventional non-steroidal anti-inflammatory drugs (NSAIDs) and compared to newer NSAIDs or different pain medication. Such abstracts could only be excluded by analyzing the abstracts themselves. Whether to proceed systematically back through the abstracts ordered by recency, or to retrieve abstracts from a random time interval, or from a window of n-years after the drug came on the market, will be a matter to be assessed empirically.

The final source of "non relevant" judgments was a problem with the judges and not with the abstracts. In Question 2 regarding dyspepsia, two out of seven abstracts were judged irrelevant because the drugs were not explicitly compared regarding dyspepsia but only regarding H. pylori, which is one of the possible causes for dyspepsia. Also abstracts retrieved for Question 7 about the effect on lipid profiles were wrongly categorised by roughly a third of the judges as not being relevant to the liver.

The experiment has shown that searching for the drugs, the basis of the comparison and studies of the publication type *comparative study* is a first step towards retrieving abstracts that can serve as answer candidates for clinical comparison questions, but it has been shown not to be sufficient to guarantee the relevance of the retrieved abstracts.

The two main problems discovered during the experiment that need to be addressed in further processing steps for the retrieved abstracts concern abstracts lacking sentences in which the drugs are directly compared to each other and the retrieval of irrelevant abstracts for well-established drugs, which are used as a reference for comparing newer drugs to, instead of containing direct comparisons of the drugs in question.

7 Conclusion and Future Work

This work introduced the task of answering clinical comparison questions and pointed out challenges in recognising and extracting their components. It also described strategies for retrieving MEDLINE® abstracts and showed that only looking for the compared entities without including the basis of the comparison is not enough to retrieve useful abstracts.

The initial experiment evaluating the relevance of retrieved abstracts for twelve clinical comparison questions revealed a number of problems that need to be taken into account for future work, especially the lack of sentences containing explicit comparisons and dealing with well-established drugs.

During the next stages, the process of identifying and extracting the elements of a comparison question as well as the process of retrieving MEDLINE® abstracts will be automated using tools from the UMLS Knowledge Sources. Features or rules will be defined to augment SemRep to deal with the problems concerning *non-scalar* comparisons and structurally different *scalar* comparison discussed in Section 3.1 to be able to automatically extract the relevant comparison components. Also, possible solutions will be researched to automatically overcome the problems of retrieving relevant abstracts identified and discussed in Section 6.

Acknowledgments

I would like to thank my advisors Bonnie Webber, Marshall Dozier and Claudia Pagliari for their helpful comments and support. I would also like to thank John Brassey for providing insight into the NLH and ATTRACT QAS.

References

Thorsten Brants. TnT – A Statistical Part-of-Speech Tagger. Available at http://www.coli.uni-saarland.de/~thorsten/publications/Brants-TR-TnT.pdf Accessed 10 August 2008.

Lacey Sue Bryant and Tim Ringrose (2005). Clinical Question Answering Services: What users want and what providers provide. Poster.

M Lee Chambliss and Jennifer Conley (1996). Answering Clinical Questions. J**ournal of Family Practice** 43: 140–144.

Dina Demner-Fushman and Jimmy Lin (2006). Answer Extraction, Semantic Clustering, and Extractive Summarization for Clinical Question Answering. **Proc. COLING/ACL 2006**: 841–848.

John W Ely, Jerome A Osheroff and Mark H Ebell (1999). Analysis of questions asked by family doctors regarding patient care. **BMJ** 319: 358–361.

John W Ely, Jerome A Osheroff, M Lee Chambliss, et al. (2005). Answering physicians' clinical questions: Obstacles and potential solutions. **Journal of the American Medical Informatics Association** 12(2): 217–224.

Marcelo Fiszman, Dina Demner-Fushman, Francois M. Lang et. al. (2007). Interpreting comparative constructions in biomedical text. **Proc. BioNLP 2007**: 137–144.

Joseph L Fleiss (1971). Measuring nominal scale agreement among many raters. **Psychological Bulletin** 76 (5): 378–382.

Carol Friedman (1989). A general computational treatment of the comparative. **Proc. ACL 1989**: 161–168.

Rodney Huddleston and Geoffrey K Pullum (eds.) 2002. *The Cambridge Grammar of the English Language*. Cambridge: Cambridge University Press.

Thomas C Rindflesch and Marcelo Fiszman (2003). The interaction of domain knowledge and linguistic structure in natural language processing: Interpreting hypernymic propositions in biomedical text. **JBI** 36(6): 462–77.

Thomas C Rindflesch, M Fiszman and Bisharah Libbus (2005). Semantic interpretation for the biomedical research literature. *Medical informatics: Knowledge management and data mining in biomedicine.* Springer, New York, NY.

David L Sackett, William M C Rosenberg, J A Muir Gray, et al. (1996). Evidence based medicine: what is is and what it isn't: It's about integrating individual clinical expertise and the best external evidence. **BMJ** 312, pp. 71–72.

Silke Scheible (2008). Annotating Superlatives. **Proc. LREC 2008:** 28–30.