Helping parents to understand rare diseases

Marina Sokolova
CHEO Research Institute and University of Ottawa
sokolova@uottawa.ca

Ilya Ioshikhes
University of Ottawa
iioschik@uottawa.ca

Hamid Poursepanj
University of Ottawa
Hpour099@uottawa.ca

Alex MacKenzie
CHEO Research Institute and University of Ottawa
mackenzie@cheo.on.ca

Abstract

Rare diseases are not that rare: worldwide, one in 12-17 people will be affected by a rare disease. Newborn screening for rare diseases has been adopted by many European and North American jurisdictions. The results of genetic testing are given to millions of families and children's guardians who often turn to the Internet to find more information about the disease. We found 42 medical forums and blogs where parents and other related adults form virtual communities to discuss the disease diagnosis, share related knowledge and seek moral support. Many people (up to 75% in some population groups) look for professional medical publications to find reliable information. How can it be made easier for these non-medical professionals to understand such texts? We suggest that recommender systems, installed on web sites of research and teaching health care organizations, can be a tool that helps parents to navigate a massive amount of available medical information. In this paper, we discuss NLP architecture of such a system. We concentrate on processing epistemic modal expressions and helping the general public to evaluate the certainty of an event.

1 Introduction

A rare disease is identified as a life-threatening or chronically debilitating condition affecting not more than 5 in 10.000 persons (Cornel et al., 2013). Accumulatively, rare diseases are not uncommon. In UK, one in 17 people will be affected at some point in their lives; this equates to more than 3.5 million people. Most of these will be children, and 30% of individuals with a rare disease will die before they are five years old. [1] In Canada, one in 12 people suffer from a rare disease, and the number of identified rare diseases identified constantly increases.[2]

A prevailing understanding is that early diagnosis and treatment may ease the burden of a disease. Thus, newborn screening for rare diseases has become a routine procedure in USA, Canada and the member states of the European Union. Nevertheless, Raffle and Gray (2007) note that screening programs can induce harm. Whereas the programs improve health status in patients by diagnosing them early and treating optimally, after the screening, parents and guardians of the newborns receive a substantial amount of new information about health of their children. The results may cause parental stress and anxiety, among other negative factors. To aid in navigation through the screening results, health care authorities organized web-based resource centers, such as a joint web site by the Newborn Screening Ontario program and Children's Hospital of Eastern Ontario[3] or the Newborn Screening web site by the National Health Service[4].

Emergence of user-friendly online technologies prompted the general public to turn to the Internet to gain more knowledge on health-related issues, a phenomenon often referred to as Dr. Google. A 2011 survey of the US population estimated that i) 59% of all adults have looked online for information about health topics such

[1] http://blogs.biomedcentral.com/bmcblog/2013/02/28/what-is-the-cost-of-rare-diseases/
[2] http://rare-diseases.ca/
[3] http://www.newbornscreening.on.ca/bins/index.asp
[4] http://www.nhs.uk/Livewell/Screening/Pages/Newbornscreening.aspx

Proceedings of the Workshop on NLP for Medicine and Biology associated with RANLP 2013, pages 24–33,
Hissar, Bulgaria, 13 September 2013.

as a specific disease or treatment, ii) 18% of adults have consulted online reviews of particular drugs or medical treatments, iii) 29% of all adults sought online health information related to somebody's else medical condition (Fox 2011; Fox 2011a). Preference in online searches related to specific health problems was previously reported by Nicholas et al. (2003).

At the same time, the general public does not consider different sources of the available health information as being equal. 75% of non-medical professionals aim their online searchers at professional medical sites and publications (McMullan, 2006). Many individuals prefer to have an access to a complex and complete information (80%), whereas some feel that the information usually accessed is too basic (45%) (ibid).
An important question arises as to how well the readers can understand the information they retrieve. Eysenbach (2003) reported that patients who sought health information online felt that Internet information can be overwhelming (31%), conflicting (76%) and confusing (27%).

The system that we are building aims to help individuals, specifically parents of newborns, to navigate and assess medical publications about rare diseases.

2 Motivation

It has been documented that many parental users of the Internet are first-time parents (Oprescu et al, 2013) and parents with young children (Balka and Butt, 2006). They might not have a hands-on experience or practice in dealing with a complex medical issue and information related to it and can be easily overwhelmed. This is especially true for parents of newborn babies which were screened positive for one of the rare diseases.

The screening happens at the very beginning of the child's life. A positive result may cause parents' insecurity and increase their uncertainty about future (Brashers, 2001). Gaining new knowledge through relevant information is one of the tools that can alleviate stress and anxiety (Brashers, 2000). It is natural that the affected parents search for information about the diagnosis and turn their attention to professional medical publications.

In the quest for knowledge, parents face a challenge of understanding a complex text that includes assessment of the relevance of the retrieved information and ability to differentiate among available options. This necessitates, among other required skills, ability to discriminate between different degrees of certainty and evaluate the likelihood score found in the text (Holmes, 1982; Morante and Sporleder, 2012).

3 Medical Forum Data

To get a notion of the public concerns and questions, we automatically extracted and manually analyzed messages posted on medical forums and blogs dedicated to newborn screening for rare diseases.

We selected 42 forums frequently visited by parents and families concerned about newborn screening and its consequences (e.g., forums.familyeducation.com,www.justmommies.com/babies/newborn,www.parentingforums.org/forum.php). We did not require research ethics review for this study as all of the data collected and used was from publically available sources. Nevertheless, we confirmed with our institutional research ethics board that no review of research on public data sets was necessary.

To find what parents think and discuss, we followed two strategies to collect data: a manual search and automated crawling of parenting forums and blogs. In our manual data collection, we used Google to find blogs, comments and medical forums that could post messages with the relevant content. Our queries were built from all the possible combinations of the following three sets of phrases:

- {Forum, Bulletin Board, Message Board, BBS, Threads}
- {Newborn, infant}
- {Genetic Predisposition Testing, Genetic Screening, Predictive Genetic Testing, Genome Sequencing}

For instance, the query {newborn genetic screening threads} is used to search for the related forums. After finding relevant blogs or threads, we downloaded the comments.

To download the comments we used Selenium API[5], Hibernate API[6], MySQL database, and Java programming language. In automatic data collection, we used Apache Nutc crawler application[7] in combination with Apache Solr search platform[8] to crawl handpicked forums.

First, we performed manual search to find forums with the related phrases mentioned above. Then we used URLs of the selected forums as a seed for Nutch. We stored the collected HTML pages in MySQL database. Next, we searched these pages with Apache Solr, which uses the cosine similarity metric to find related page to the search query:

$$\text{similarity} = \cos(\theta) = \frac{A \cdot B}{\|A\|\|B\|} = \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n} (A_i)^2} \times \sqrt{\sum_{i=1}^{n} (B_i)^2}}$$

where A and B are the term frequency vectors of the query and a document. We marked page as relevant if it contained any three-phrase combination built from three sets of phrases mentioned earlier in this section.

After finding pages with relevant comments, their content was downloaded and analyzed independently by two authors. We found that parents and other involved individuals are often concerned about the following issues:

- Prevalence and severity of the disease
- Available treatment and the effects of an early treatment
- A possible course of the disease
- Reliable tests
- Health care facilities
- Medications and their side effects
- Future use of the results

Table 1 lists some examples of messages. We keep all the original spelling, punctuation, and grammar.

Table 1: Online messages on newborn screening

Concern	Message
Prevalence and severity	I was 39 when my son was born. 1:810 for downs and 1:10,000 for tri 18/13. The risk for ANY chromosomal issue is 1:80 for someone age 39.
Available treatment and the effects of an early treatment	she said if he starts to get congestion then do a breathing treatment and call next day if he is not doing better. I was confused about this as well as I thought he would need those treatments immediately....
A possible course of the disease	nothing is written in stone for CF these days. Docs told my parents I wouldn't live past 21. I'm 27yo, working on a PhD, and this past spring I finished my second Half-Ironman Triathlon.
Reliable tests	I'm concerned because of the number of false-positives I read about, and further testing to eliminate that worry have a small chance of causing death to a perfectly healthy baby.
Health care facilities	best to do them in a cf[9] center cause it all depends on the lab you do the test in.
Medications and their side effects	I have a 6 month old son that was born deaf. He just recieved a baha hearing aid, and it has done some good, but not as much as we had hoped for
Future use of the results	And you have no problem with this government owning their genetic code, potentially knowing illness, disabilities, strengths, weaknesses and potential? A trusting soul you are indeed.

[5] http://docs.seleniumhq.org/, accessed: July. 2, 2013
[6] http://www.hibernate.org/, accessed: July. 2, 2013

[7] http://nutch.apache.org/#What+is+Apache+Nutch%3F, accessed: July. 2, 2013
[8] http://lucene.apache.org/solr/ , accessed: July. 2, 2013

[9] Cystic fibrosis - an autosomal recessive genetic disorder that affects most critically the lungs, and also the pancreas, liver, and intestine.

4 Epistemic expressions

Our current project focuses on disambiguating epistemic modal expressions. Previously, several studies connected the use of modal verbs (can, might, should), amplifiers (certainly, definitely) with the level of expectation of an event (Henriksson and Velupillai, 2010; Sokolova and Lapalme, 2011).

However, these studies did not categorize epistemic expressions based on their strength of conviction in the event happening. We categorize the strength of conviction by assigning six categories: *impossible*, *improbable*, *uncertain*, *possible*, *probable*, and *certain* (Horn, 1989; Sokolova et al, 2010). Table 2 shows the categories and expressions.

Table 2: Examples of epistemic expressions

Epistemic category	Expressions
Impossible	Never happens
Improbable	Hardly expected
Uncertain	We unsure
Possible	Perhaps
Probable	There is a certain risk
Certain	We always see

The suggested six categories add three negative categories (*impossible*, *improbable*, *uncertain*) to common positive happenstance categories (*possible*, *probable*, *certain*) (R. Saurı´ and J. Pustejovsky, 2009).

Language expressions corresponding to the categories contain extensional modifiers, i.e. modifiers of degree and happenstance (Sokolova and Lapalme, 2011). The modifiers can be modal verbs (can, must, would), adverbs (likely, mostly), adjectives (common, rare) and quantifying pronouns (every, none). These expressions should be disambiguated in the context of a sentence or a clause.

Below we categorize a few expressions found in articles on rare diseases:

- Impossible: such off-target gene modulating effects are currently impossible to predict
- Improbable: in part out of concern that recruitment of eligible subjects with SMA Type I would be difficult because of their high level of inter-current illness and mortality in childhood
- Uncertain: these studies could not rule out an additional contribution resulting from restoration of SMN levels in muscle
- Possible: raising the possibility that intrinsic responses to low levels of SMN in skeletal muscle may also contribute directly to SMA pathogenesis
- Probable: Studies have shown that restoring SMN protein levels in neurons can significantly ameliorate disease progression
- Certain: A neighboring nearly-identical copy of this gene, SMN2, is invariably present in individuals with SMA

Our next step was to assess the presence of epistemic expressions in articles on rare diseases.

5 Empirical evidence

For our preliminary study we decided on a group of articles dedicated to spinal muscular atrophy (SMA), a neurodegenerative disease affecting 1 in 11000 newborns world- wide (Farooq et al, 2013). We selected six full-length research articles. The articles were published in Orphanet Journal of Rare Diseases, Neurodegenerative Diseases (an open source book), Journal of Clinical Investigation**,** Plos One, and Human Molecular Genetics (2 articles).

The articles' content covers most of parent concerns (see Table 1): a) severity of the disease, b) available treatment and the effects of an early treatment, c) a possible course of the disease, reliability of tests, d) medications. Thus, the article selection allows the empirical results to be representative of the information parents will find. At the same time, we cover language expressions belonging to different authors.

Three articles were written by a team of leading researchers in SMA; these authors have written papers that receive a high rank in SMA search through Google Scholar; hence, there is a high probability that parents looking for the SMA information will first encounter papers published by this team (Articles A). Three other articles were written by other teams working on SMA research (Articles B). We report the descriptive statistics in Table 3; *vocabulary* signifies different words in the text.

Table 3: Vocabulary richness of the six articles; *d.leg*. – vocabulary with occur. = 1, *h.leg.* – vocabulary with occur. = 2, occ > 5 – vocabulary that occur 6 and more times.

Articles A					
#	**words**	**vocab.**	***d. leg.***	***h. leg.***	**Occ >5**
1a	6894	1661	1008	254	159
2a	4492	1374	827	242	121
3a	6629	1506	871	246	185
Articles B					
#	**words**	**vocab.**	***d. leg.***	***h. leg.***	**Occ >5**
1b	10731	2658	1344	441	317
2b	6009	1557	896	246	173
3b	5094	1125	598	191	157

We looked for epistemic expressions related to the main concerns found on medical forums (see Table 1). For the information retrieval, we built N-gram models of each article (N = 1, …, 4). We tokenized data by splitting along spaces and punctuation marks. We kept the original capitalization to preserve the beginning of sentences. We used combinations of the seed words to find N-grams with the epistemic meaning. Table 4 lists examples of the seed words.

Table 4: Examples of words used in search of epistemic expressions.

Part-of-speech	**Examples**
Adverbs	Possibly, perhaps
Adjectives	Impossible
Modal verbs	Can, may, should
Negations	No, not
Nouns	absence, presence
Quantifying pronouns	Every, none

To avoid counting the same expression multiple times, we filtered out those bi-grams which first word overlaps with the second word of another epistemic bi-gram. From the list of tri-grams, we filtered out those which first word overlaps with the third word of another epistemic tri-gram. For instance, we filtered out not be as a possible extension of may not.

Manual analysis showed that the most frequent bi- and tri-grams expressed negative and positive conviction, supporting the proposed expansion of epistemic categories by negative *impossible*, *improbable* and *uncertain.* Table 5 lists the most frequent epistemic bi- and tri-grams found in the articles.

Table 5: Five most frequent not overlapping epistemic bi- and tri-grams per article.

Articles A		
Article1a	Bi-grams	may not, may be, can be, will be, where possible
	Tri-grams	may not be, the hope is, would have a, majority of these, and where possible
Article 2a	Bi-grams	which can, SMA can, and can, must be, may have
	Tri-grams	which can cross, no cure for, SMA can be, SMA is primarily, which may have
Article 3a	Bi-grams	potential treatment, such promising, promise for, of approximately, suggest that
	Tri-grams	potential therapeutic compounds, potential treatment strategy, such promising agent, found to be, There could be
Articles B		
Article 1b	Bi-grams	can be, able to, will be, would be, could be
	Tri-grams	the potential to, the false discovery, can be used, may not be, there were no
Article 2b	Bi-grams	absence of, implicated in, as expected, potential to, the possibility
	Tri-grams	absence of any, confirmed in a, raising the possibility, potential to act, supported by several
Article 3b	Bi-grams	a putative, probably carry, could be, should be, would be
	Tri-grams	result not shown, putative gene conversion, affected children could, observations should be, This change would

We hypothesized that the use of the expressions closely relates to the content of the article. For example, articles reporting on clinical trials, treatments, medications may have a high frequency of epistemic expressions as due to necessity of drawing conclusions and implications for future patients. We list the topics of the six papers and frequencies of the top epistemic bi- and tri-grams in Table 6.

Table 6: Article topics and frequencies of the epistemic expressions ($x10^{-3}$)

Articles A			
#	Topics	bigrams	trigrams
1a	disease therapy, preclinical drug development, generalizable screening methods	22.4	14.0
2a	Classification, diagnosis, background for SMA	5.0	2.9
3a	PRL treatment in mice, potential therapeutic compounds	2.1	1.6
1b	Identification of novel candidate biomarkers associated with disease severity in SMA	3.2	1.1
2b	intrinsic pathology of skeletal muscle, novel biomarkers in SMA	6.3	2.0
3b	molecular analysis of the SMN and NAIP Genes	4.3	2.3

Looking at the topics of the papers, we can see that preclinical drug development corresponds to the largest frequency of the epistemic expressions in the text.

6 Parent Advisor

In the medical domain, uncertainty and misunderstanding of information can imperil lives and incur significant costs on health care systems (McCoy et al, 2012). Although a thirst for medical knowledge among non-medical readers has been documented (see Section 1), there are not many developed NLP tools and methods that help such readers to understand medical texts.

Although epistemic disambiguation is important for text understanding, other text analysis tasks are essential in order to build an effective system that helps parents to understand medical publications related to rare diseases. Hence, the proposed system name is the Parent Advisor.

We suggest that such the Parent Advisor is organized as a pipeline of NLP and Text Mining tools, each serving a special purpose (Figure 1):

1. social media analyzer, to identify concept shift in parents' concerns;
2. article content classifier, to select relevant articles for further analysis;
3. relevant article ranker, to evaluate the usefulness of the selected articles;
4. factual information extractor, to retrieve information related to parents' concerns
5. epistemic disambiguator, to assess the retrieved factual information.
6. the output ranker, which ranks the factual information according to the assigned epistemic categories.

We also suggest that human participation should be incorporated in the system functioning. To support the actuality of text analysis, parents can be surveyed and polled either on a regular basis or in relation to medical and health-related events (e.g., a new discovery, a proposed change in health care) (Fox, 2011). Medical professionals should be involved in the article ranking, to ensure the quality of the selected publications. Additionally to standard text annotation, a team of communication and medical professionals can assist in the factual information extraction and epistemic disambiguation (Scott et al, 2012).

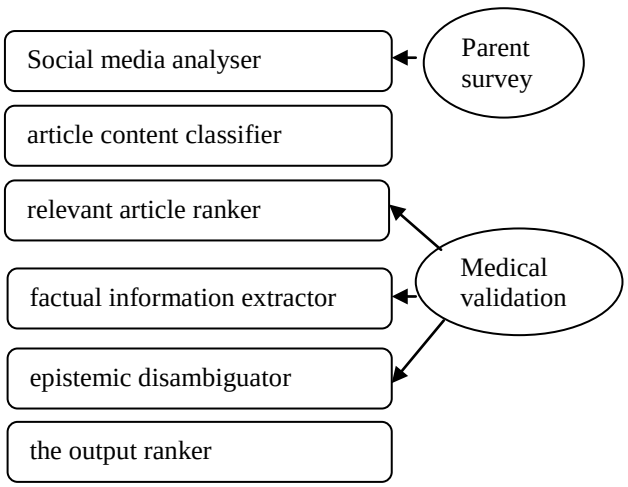

Figure 1: the Parent Advisor system.

For a given query, the system will release the information ranked accordingly to the epistemic categories. For example, the information deemed certain will be ranked higher than the information deemed probable or possible (Table 7). When releasing the information marked with negative categories, we plan to label it with "Caution".

Table 7: A possible output for the query " orphan diseases, protein, therapy".

#	Extracted text	Annotation (not shown to parents)
1	Pharmacological chaperones stabilize the folding of mutant proteins and allow for correct trafficking of the enzyme	*Certain*
2	mRNA serves as a valid proxy for protein level more often than not.	*Probable*
Caution	[the pharmacologic upregulation of gene activity and mRNA level] will not work if the mutated protein has any dominant negative effect.	*Impossible*
Caution	even if a protein:RNA correlation is observed in vitro, a given transcript response detected in cell culture may not hold true for a whole organism.	*Improbable*

7 Related Work

The semantic analysis of biomedical and clinical texts mainly focuses on identification and disambiguation of medical terms and events (Cohen et al, 2011; Demner-Fushman et al, 2010; Savova et al, 2011) including temporal characteristics of events (Boytcheva et al, 2012). Emergence of electronic health records enabled studies of epistemic expressions, often linking them with the diagnosis of a patient (McCoy et al, 2012; S. Velupillai' 2010). Expressions of certainty in the medical publications, however, are not well studied, although it is natural to expect a medical publication to contain epistemic expressions: while presenting factual information, a publication also conveys the authors' inference from the facts and conviction in the event happenstance.

Categorizing text into speculative and non-speculative parts partially addresses the problem (Szarvas, 2008; Sanchez et al, 2010) as such division only differentiates the *certain* (aka non-speculative) category from other epistemic categories. We, however, want to analyze language expressions of several epistemic categories.

Vagueness in clinical records was studied by Emanuel and Emanuel (1989) and more recently by Hyland (2006) and Scott et al (2012). These studies concentrate on the use of happenstance modifiers (Sokolova and Lapalme, 2011), also known as linguistic hedges (e.g., *possible, probably*, *few*, *consistent with*). The goal is to develop an automated analysis of diagnosis and symptom information extracted from electronic patient records (Scott et al, 2012). Although the task is similar to our goal, we plan to work with information extracted from medical publications.

A more complex approach would be to introduce a pragmatic component into the epistemic analysis of medical publications. This approach differentiates between a hypothesis, accepted knowledge, and new experimental knowledge (Nawaz et al, 2010). Such categorization may be useful in the recommender system designed for the general public. In future, we plan to work on the hypothesis - experimental knowledge division.

Our work also closely relates to text understanding and interpretation (Bos, 2011). With the development of Internet search engines, text understanding and interpretation mainly focused on retrieval of texts relevant to the query. A few systems develop a more advanced and deep text exploration which interprets text on demand from the system users (Dunne et al, 2012). The systems are field-specific, often built on ontology, and are designed to help professionals working in the field (Dunne et al, 2012; Wimalasuriya and Dou, 2010). An advanced, semantically based information retrieval is performed by question-answering systems. In medicine, such systems assist clinicians to find clinically-relevant information (Cao et al, 2011). Our goal, in con-

trast, is to build a system that helps the general public to understand professional medical text.

Note that during the related work analysis, we could not find published studies that relate parent concerns to social media to rare disease information.

8 Future Work

Our immediate future work will focus on epistemic annotation of a large collection of SMA articles. What can be considered a sufficient size of the annotated corpus is an open question in BioNLP: the physical chemistry and biochemistry Core Scientific Concepts corpus has 265 articles (Liakata, 2010), the bioinformatics' Bioscope corpus has 9 articles and 1273 abstracts (Vincze et al, 2008).

We suggest that the number of annotated articles can be linked to the annual rate of SMA publications. For example, the Google Scholar search for "spinal muscular atrophy" [10] retrieved 278 articles [11] published in 2012 – 2013. A similar PubMed search [12] retrieved 222 articles [13]. Hence, we aim to annotate 150 - 200 SMA articles and abstracts.

We want the article profiles be representative of the concerns expressed on the medical forums (listed in Table 1). For example, after adding the term "treatment" to both searchers we retrieved 77 articles through Google Scholar and 99 articles through PubMed, while substituting "treatment" by "drugs" we retrieved 45 articles and 7 articles respectively. Thus, our annotated data should include 60-80 treatment-oriented and 10-30 drug-oriented articles and abstracts.

The article and abstract selection is another open question in the article annotation. Our criterion is the corpus compliance with the expected retrieval results of the Web search. Hence, for each topic, we will select articles based on their relevance.

To ensure that the corpus is consistent with the public concerns, we will continue to analyze medical forums in order to keep updates on the general public response on newborn screening for rare diseases.

9 Conclusions

We presented a preliminary work on the system which we call Parent Advisor. Parent Advisor can help parents to understand medical publications related to rare diseases. The topic of rare diseases became a subject of many discussions, as more and more jurisdictions adopted a policy of newborn screening for rare diseases.

To make our future system actual and useful for parents, we have studied messages posted on 42 medical forums. These forums are frequently visited by parents and families who have questions related to newborn screening for rare diseases. We identified several issues that concern the general public most (prevalence and severity of the disease, available treatment and the effects of an early treatment, a possible course of the disease, reliable tests, health care facilities, medications and their side effects, future use of the results).

We chose epistemic disambiguation as the focus of our first sub-project in building Parent Advisor. We identified six epistemic categories (*impossible*, *improbable*, *uncertain*, *possible*, *probable*, *certain*), thus expanding a usual range of positive categories (*possible*, *probable* *certain*) by negative categories (*impossible*, *improbable*, *uncertain*).

In this paper, we also outlined the architecture of the system, sketched human-system collaboration desirable for the system to be effective, and presented a detailed plan for future work.

Acknowledgments

This work was in part supported by Natural Sciences and Engineering Research Council of Canada Discovery Grant. The authors thank anonymous reviewers for helpful comments.

[10] http://scholar.google.ca/scholar?hl=en&as_sdt=1,5&as_vis=1&q=%22spinal+muscular+atrophy%22&scisbd=1, accessed Aug 13, 2013.
[11] All languages.
[12] http://www.ncbi.nlm.nih.gov/pubmed, accessed Aug 13, 2013.
[13] English only.

References

E. Balka and A. Butt. 2006. Information Seeking Practices for Youth, Parents and Seniors. *Report.* www.sfu.ca/act4hlth/

Bos, J. 2011. A Survey of Computational Semantics: Representation, Inference and Knowledge in Wide-Coverage Text Understanding. *Language and Linguistics Compass*, 5: 336–366.

S. Boytcheva, G. Angelova, I. Nikolova. 2012. Automatic Analysis of Patient History Episodes in Bulgarian Hospital Discharge Letters. *Proceedings of EACL*, 77-81

D. Brashers. 2001. Communication and Uncertainty Management. *Journal of Communication*, 51(3):477-497

D. Brashers, J. Neidig, S. Haas, L. Dobbs, L. Cardillo, J. Russell. 2000. Communication in the management of uncertainty: The case of persons living with HIV or AIDS. *Commun Monogr* , 67(1):63-84

Y. Cao, F. Liu, P. Simpson, L. Antieau, A. Bennett, J. Cimino, J. Ely, H. Yu. 2011. AskHERMES: An online question answering system for complex clinical questions, *Journal of Biomedical Informatics*, 44(2): 277-288.

K. Cohen, K. Verspoor, H. Johnson, C. Roeder, P. Ogren, W. Baumgartner Jr, E. White, H. Tipney, and L. Hunter. 2011. High-Precision Biological Event Extraction: Effects of System and Of Data. *Computational Intelligence*, 27: 681–701

M. Cornel, T. Rigter, S. Weinreich, P. Burgard, G. Hoffmann, M. Lindner, G. Loeber, K. Rupp, D. Taruscio and L. Vittozzi. 2013. A framework to start the debate on neonatal screening policies in the EU: an Expert Opinion Document, *European Journal of Human Genetics*, 1-6.

D. Demner-Fushman, J. Mork, S. Shooshan, A. Aronson. 2010. UMLS Content Views Appropriate for NLP Processing of the Biomedical Literature vs. Clinical Text. *Journal of Biomedical Informatics*, 43(4): 587–594.

Dunne, C., Shneiderman, B., Gove, R., Klavans, J. and Dorr, B. 2012. Rapid understanding of scientific paper collections: Integrating statistics, text analytics, and visualization. *J. Am. Soc. Inf. Sci.*, 63: 2351–2369

L. Emanuel and E. Emanuel. 1989. The medical directive: A new comprehensive advance care document. *Journal of the American Medical Association*, 261(22): 3288 – 3293.

G. Eysenbach. 2003. The impact of the Internet on cancer outcomes. *CA Cancer J Clin*, 53:356–71.

F. Farooq, F. Abadı´a-Molina, D. MacKenzie, J. Hadwen, F. Shamim, S. O'Reilly, M. Holcik, and A. MacKenzie. 2013. Celecoxib increases SMN and survival in a severe spinal muscular atrophy mouse model via p38 pathway activation. *Human Molecular Genetics*, 1–10.

S. Fox. 2011. *The Social Life of Health Information*. Pew Research Center's Internet & American Life Project,http://pewinternet.org/Reports/2011/Social-Life-of-Health-Info.aspx

S. Fox. 2011a. *Survey Questions.* Pew Research Center's Internet & American Life Project, http://pewinternet.org/Reports/2011/Social-Life-of-Health-Info.aspx

A. Henriksson and S. Velupillai. 2010. Levels of certainty in knowledge-intensive corpora: An initial annotation study. *Proceedings of the Workshop on Negation and Speculation in Natural Language Processing*, pages 41–45

J. Holmes. 1982. Expressing Doubt and Certainty in English. *RELC Journal*, 13:9-28.

L. Horn. 1989. *A Natural History of Negation*. The University of Chicago Press.

K. Hyland. 2006. Medical discourse: hedges. *Encyclopedia of Language and Linguistics*, p.p. 694- 697.

M. Liakata. 2010. Zones of conceptualisation in scientific papers: a window to negative and speculative statements. *Proceedings of the Workshop on Negation and Speculation in NLP*, 1–4.

W. McCoy, C. Alm, C. Calvelli, J. Pelz, P. Shi, A. Haake. 2012. Linking Uncertainty in Physicians' Narratives to Diagnostic Correctness. *Proceedings of the ACL-2012 Workshop on Extra-Propositional Aspects of Meaning in Computational Linguistics*

M. McMullan. 2006. Patients using the Internet to obtain health information: How this affects the patient–health professional relationship. *Patient Education and Counseling*, 63:24–28, Elsevier.

R. Morante and C. Sporleder. 2012. Modality and Negation: An Introduction to the Special Issue. *Computational Linguistics*, 38(2): 224-260.

R. Nawaz, P. Thompson, S. Ananiadou. 2010. Evaluating a Meta-Knowledge Annotation Scheme for Bio-Events, *Proceedings of the Workshop on Negation and Speculation in NLP*, pages 69–77.

D. Nicholas , P. Huntington, B. Gunter, C. Russell, R. Withey. 2003. The British and their use of the web for health information and advice: a survey. *Aslib Proc*, 55:261–76.

F. Oprescu, S. Campo, J. Lowe, J. Andsager, J. Morcuende. 2013. Online Information Exchanges for Parents of Children with a Rare Health Condition: Key Findings From an Online Support Community. *Journal of Medical Internet Research*, 15(1):e16

A.Raffle, M. Gray. 2007. *Screening. Evidence and Practice*. Oxford: Oxford University Press.

L. Sanchez, B. Li, C. Vogel. 2010. Exploiting CCG Structures with Tree Kernels for Speculation Detection. *Proceedings of CoNLL: Shared Task*, 126–131.

R. Saurı´ and J. Pustejovsky. 2009. FactBank: a corpus annotated with event factuality. Language Resources and Evaluation, 43:227–268.

G. Savova, W. Chapman, J. Zheng, R. Crowley. 2011. Anaphoric relations in the clinical narrative: corpus creation. *Journal of American Medical Informatics Association*, 18(4): 459-465.

D. Scott, R. Barone, B. Koeling. 2012. Corpus annotation as a scientific task. *Proceedings of LREC'2012*, p.p. 1481 – 1485.

M. Sokolova, K. El Emam, S. Chowdhury, E. Neri, S. Rose, E. Jonker. 2010. Evaluation of Rare Event Detection, *Advances in Artificial Intelligence 23*, pp. 379–383

M. Sokolova and G. Lapalme. 2011. Learing opinions in user-generated Web content", *Journal of Natural Language Engineering*, Cambridge University Press, 17(4): 541–567

S. Velupillai. 2010. Towards A Better Understanding of Uncertainties and Speculations in Swedish Clinical Text – Analysis of an Initial Annotation Trial. *Proceedings of the Workshop on Negation and Speculation in NLP*, 14–22.

V. Vincze, G. Szarvas, R. Farkas, G. Mora, and J. Csirik. 2008. The BioScope Corpus: Biomedical Texts Annotated for Uncertainty, Negation and their Scopes. *BMC Bioinformatics*, 9(Suppl 11):S9.

D. Wimalasuriya and D. Dou. 2010. Ontology-based information extraction: An introduction and a survey of current approaches. *Journal of Information Science*, 36(3):306 – 323