Negation and Speculation in Natural Language Processing (NeSp-NLP 2010)

Proceedings of the Workshop

10 July 2010
Uppsala, Sweden

Roser Morante and Caroline Sporleder (eds.)

Proceedings of the Workshop on
Negation and Speculation in Natural Language Processing
(NeSp-NLP 2010)

July 2010

ISBN: 9789057282669
EAN: 9789057282669

University of Antwerp
Prinsstraat 13
B-2000 Antwerp
Belgium
Tel: +32(0)3 265 41 11
Fax: +32(0)3 265 44 20
http://www.ua.ac.be

These proceedings were compiled with the ACLPUB package.

Organizers:

Roser Morante - University of Antwerp (Belgium)
Caroline Sporleder - Saarland University (Germany)

Program Committee:

Timothy Baldwin - University Melbourne (Australia)
Xavier Carreras - Technical University of Catalonia (Spain)
Wendy W. Chapman - University of Pittsburgh (USA)
Kevin B. Cohen - University of Colorado (USA)
Walter Daelemans - University of Antwerp (Belgium)
Guy De Pauw - University of Antwerp (Belgium)
Bonnie Dorr - University of Maryland (USA)
Roxana Girju - University of Illinois at Urbana-Champaign (USA)
Iris Hendrickx - University of Lisbon (Portugal)
Veronique Hoste - University College Ghent (Belgium)
Halil Kilicoglu - Concordia University (Canada)
Martin Krallinger - CNIO (Spain)
Lori Levin - Carnegie Mellon University (USA)
Maria Liakata - Aberystwyth University, European Bioinformatics Institute (UK)
Lluis Màrquez - Technical University of Catalonia (Spain)
Erwin Marsi - Tilburg University (The Netherlands)
Arzucan Özgür - University of Michigan (USA)
Manfred Pinkal - Saarland University (Germay)
Sampo Pyysalo - University of Tokyo (Japan)
Owen Rambow - Columbia University (USA)
Josef Ruppenhofer - Saarland University (Germany)
Roser Saurí - Barcelona Media Innovation Center (Spain)
Khalil Sima'an - University of Amsterdam (The Netherlands)
Mihai Surdeanu - Stanford University (USA)
Antal van den Bosch - Tilburg University (The Netherlands)
Michael Wiegand - Saarland University (Germany)

Invited Presentations:

Ed Hovy - ISI, University of Southern California (USA)
Martin Krallinger - CNIO (Spain)
Maria Liakata - Aberystwyth University, European Bioinformatics Institute (UK)
Raheel Nawaz, Paul Thompson, Sophia Ananiadou - NaCTeM (UK)
Veronika Vincze - BioScope group, University of Szeged (Hungary)

Institutions that organize the workshop:

ACL SIGs that endorse the workshop:

Website:

http://www.clips.ua.ac.be/NeSpNLP2010/

Sponsors:

GOA project Biograph from the University of Antwerp
Cluster of Excellence, MMCI, Saarland University

Introduction

These proceedings contain the papers and invited talks presented at the Workshop on Negation and Speculation in Natural Language Processing (NeSp-NLP 2010) that was held on the 10th of July, 2010 in Uppsala, Sweden. The program consisted of five invited talks, seven presentations of long papers and two of short papers.

When we thought of organising this workshop, we aimed at bringing together researchers working on negation and speculation from any area related to computational language learning and processing. Specific goals were to describe the lexical aspects of negation and speculation, to define how the semantics of these phenomena can be modelled for computational purposes, to explore techniques aimed at learning the factuality of an statement, and to analyse how the treatment of these phenomena affects the efficiency of Natural Language Processing (NLP) applications.

Negation and speculation are two linguistic phenomena involved in deep understanding of text. They are resources used to express the factuality of statements, which indicates to which extent a statement is or is not a fact. Negation turns an affirmative statement into negative (it rains/it does not rain). Speculation is used to express levels of certainty (it might rain/apparently, it will rain/ it is likely to rain/it is not clear whether it will rain/we suspect that it will rain). We knew that negation and speculation (or modality) have been extensively studied from a theoretical perspective. Furthermore, we also believed that there was enough interest on these topics among the NLP community and that there was enough research going on, so as to organise a topical workshop, the first of its kind as far as we know.

We cannot be exhaustive here about all the NLP related work on these topics that has been published before the workshop. We apologise for mentioning only some references as examples of research that is being carried out, which motivated our decision of organising a workshop. As recent references, the BioScope corpus has been annotated with negation and speculation cues and their scope (Vincze et al. 2009); events in the FactBank corpus (Saurí and Pustejovsky 2009) have been annotated with factuality information; the CoNLL Shared Task 2010 (Farkas et al. 2010) focused on *Learning to detect hedges and their scope in natural language text*. The biomedical text mining community has produced tools to process negation, like Context (Harkema et al. 2009), and negation has also received attention from researchers working on sentiment analysis (Wilson et al. 2009 and work cited in Wiegand et al. 2010).

We proposed the following topics in the call for papers of the workshop:

- Lexical aspects of negation and speculation
- Linguistic resources with information about negation and speculation: corpora, dictionaries, lexical databases
- Descriptive analysis of negation and speculation cues
- Negation and speculation across domains and genres
- Negation and speculation in biomedical texts and biomedical text mining
- Handling negation and speculation in NLP: dialogue systems, sentiment analysis, text mining, textual entailment, information extraction, machine translation, paraphrasing

- Learning the scope of negation and speculation cues
- Interaction of negation and speculation for evaluating the factuality of an statement
- Corpora annotation: guidelines, bootstrapping techniques, quality assessment
- Modelling factuality for computational purposes
- Algorithms to learn negation and speculation
- Structured prediction of negation and speculation
- Joint learning of negation and speculation
- Inference of factual knowledge

Although we did not receive submissions addressing all the proposed topics, the fact that we received submissions addressing some of them makes us consider that the main goal of the workshop was achieved, and that there is a growing interest in processing negation and speculation within several NLP subareas. From the nine accepted papers, six report research on biomedical texts, four of which are related to either manual or automatic annotation of corpora, one to automatically identifying negated biomedical events, and one to evaluating whether identifying negation and speculation helps in classifying medical reports. Two papers deal with negation in sentiment analysis, one focuses on automatically learning the scope and another surveys the role of negation in sentiment analysis. One paper reports research on the relation between positive and negative pairs in textual entailment.

Four of the five invited presentations are from the biomedical domain. Maria Liakata presents an annotation scheme for annotating full papers with zones of conceptualisation levels to identify the core components that constitute a scientic investigation. Veronika Vincze presents the difculties encountered during annotation process of the BioScope corpus. Martin Krallinger elaborates on the importance of negations and experimental qualifiers to extract information from biomedical literature, and Raheel Nawaz, Paul Thompson, and Sophia Ananiadou discuss the evaluation of a meta-knowledge annotation scheme for bio-events. Finally, Ed Hovy, invites us to consider Distributional Semantics as a model for richer and more semantics-oriented statistics-based NLP. He presents a specic model of Distributional Semantics, and explores the possibilities for accommodating the phenomena of negation and modality.

We would like to thank the authors of the papers for their interesting contributions, the members of the program committee for their insightful reviews, and the presenters of invited talks for accepting the invitation to give a talk at the workshop and share their work. We are grateful to Walter Daelemans for encouraging us to organise the workshop. The workshop would not have been possible without their help. We appreciate very much the knowledge, time, and effort that they invested in the workshop. We are also thankful to the University of Antwerp and Saarland University for their institutional support and to the SIGs that endorsed the workshop. We sincerely hope that in the future the NLP community will benefit from the findings made by researchers working on negation, speculation and other phenomena involved in determining the factuality of an event.

Roser Morante and Caroline Sporleder
July 2010

References

R. Farkas, V. Vincze, G. Móra, G. Szarvas, and J. Csirik (2010) The CoNLL 2010 shared task: Learning to detect hedges and their scope in natural language text. In *Proceedings of the CoNLL 2010 Shared Task*.

J. Harkema, J.N. Dowling, T. Thornblade, and W. W. Chapman (2009) ConText: An algorithm for determining negation, experiencer, and temporal status from clinical reports. *Journal of Biomedical Informatics* 42(5), 839-851.

R. Saurí and J. Pustejovsky (2009) FactBank: A Corpus Annotated with Event Factuality. *Language Resources and Evaluation*, 43(3):227-268.

V. Vincze, G. Szarvas, R. Farkas, G. Móra, and J. Csirik (2008) The BioScope corpus: biomedical texts annotated for uncertainty, negation and their scopes. *BMC Bioinformatics*, 9(Suppl 11):S9.

T. Wilson, J. Wiebe, and P. Hoffman (2009) Recognizing contextual polarity: An exploration for phrase-level analysis. *Computational Linguistics* 35:3.

M. Wiegand, A. Balahur, B. Roth, D. Klakow, and A. Montoyo (2010) A Survey on the role of negation in sentiment analysis. In this volume.

Program NeSp-NLP 2010

08.45 - 9:00	Openning
9:00 - 9:35	*Zones of conceptualisation in scientific papers: a window to negative and speculative statements* Maria Liakata (Invited presentation)
9:35 - 10:00	*Creating and evaluating a consensus for negated and speculative words in a Swedish clinical corpus* Hercules Dalianis and Maria Skeppstedt
10:00 - 10:25	*Towards a better understanding of uncertainties and speculations in Swedish clinical text Analysis of an initial annotation trial* Sumithra Velupillai
10:25 - 10:40	*Does negation really matter?* Ira Goldstein and Özlem Uzuner
10:40 - 11:10	Coffee break
11:10 - 11:45	*Speculation and negation annotation in natural language texts: what the case of BioScope might (not) reveal* Veronika Vincze (Invited presentation)
11:45 - 12:10	*Automatic annotation of speculation in biomedical texts: new perspectives and large-scale evaluation* Julien Desclés, Olfa Makkaoui, and Taouise Hacène
12:10 - 12:25	*Levels of certainty in knowledge-intensive corpora: an initial annotation study* Aron Henriksson and Sumithra Velupillai
12:25 - 12:50	*Importance of negations and experimental qualifiers in biomedical literature* Martin Krallinger (Invited presentation)
12:50 - 13.50	Lunch
13:50 - 14:50	*Negation and modality in distributional semantics* Ed Hovy (Invited presentation)
14:50 - 15:15	*What's great and what's not: Learning to classify the scope of negation for improved sentiment analysis* Isaac Councill, Ryan McDonald, and Leonid Velikovich
15:15 - 15:40	*A survey on the role of negation in sentiment analysis* Michael Wiegand, Alexandra Balahur, Benjamin Roth, Dietrich Klakow, and Andrés Montoyo
15.40 - 16.10	Coffee break
16:10 - 16:45	*Evaluating a meta-knowledge annotation scheme for bio-events* Raheel Nawaz, Paul Thompson, and Sophia Ananiadou (Invited presentation)
16:45 - 17:10	*Using SVMs with the command relation features to identify negated events in biomedical literature* Farzaneh Sarafraz and Goran Nenadic
17:10 - 17:35	*Contradiction-focused qualitative evaluation of textual entailment* Bernardo Magnini and Elena Cabrio
17:35 - 18.00	*Discussion*

Table of Contents

Zones of conceptualisation in scientific papers: a window to negative and speculative statements

Maria Liakata
Department of Computing Science, Aberystwyth University
European Bioinformatics Institute, Cambridge
liakata@ebi.ac.uk

Abstract

In view of the increasing need to facilitate processing the content of scientific papers, we present an annotation scheme for annotating full papers with zones of conceptualisation, reflecting the information structure and knowledge types which constitute a scientific investigation. The latter are the Core Scientific Concepts (CoreSCs) and include Hypothesis, Motivation, Goal, Object, Background, Method, Experiment, Model, Observation, Result and Conclusion. The CoreSC scheme has been used to annotate a corpus of 265 full papers in physical chemistry and biochemistry and we are currently automating the recognition of CoreSCs in papers. We discuss how the CoreSC scheme relates to other views of scientific papers and indeed how the former could be used to help identify negation and speculation in scientific texts.

1 Introduction

The recent surge in the numbers of papers produced, especially in the biosciences, has highlighted the need for automatic processing methods. Work by [Lin (2009)] has shown that methods such as information retrieval are more effective if zones of interest are specified within the papers. Various corpora and annotation schemes have been proposed for designating a variety of linguistic phenomena permeating scientific papers, including negation, hedges, dependencies and semantic relations [Vincze et al. (2008); Pyysalo et al. (2007); Medlock and Briscoe (2007); McIntosh and Curran (2009)]. Other schemes follow the argumentation and citation flow within papers [Teufel et al. (2009); Teufel and Siddharthan (2007)] or indeed a combination of some of the above along multiple dimensions [Shatkay et al. (2008)].

In the following we present the CoreSC annotation scheme and a corpus with CoreSC annotations. The CoreSC scheme is used at the sentence level to identify the core components that constitute a scientific investigation. We discuss how the CoreSC scheme relates to other annotation schemes representing alternate views of scientific papers and how CoreSCs could be used to guide the identification of negation and speculation.

2 The CoreSC scheme

The CoreSC annotation scheme adopts the view that a scientific paper is the human-readable representation of a scientific investigation and therefore seeks to mark the components of a scientific investigation as expressed in the text. CoreSC is ontology-motivated and originates from the CISP meta-data [Soldatova and Liakata (2007)], a subset of classes from EXPO [Soldatova and King (2006)], an ontology for the description of scientific investigations. CISP consists of the concepts: `Motivation`, `Goal`, `Object`, `Method`, `Experiment`, `Observation`, `Result` and `Conclusion`, which were validated using an on-line survey as constituting the indispensable set of concepts necessary for the description of a scientific investigation. CoreSC implements these as well as `Hypothesis`, `Model` and `Background`, as a sentence-based annotation scheme for 3-layered annotation. The first layer pertains to the previously mentioned 11 categories, the second layer is for the annotation of properties of the concepts (e.g. "New", "Old") and the third layer caters for identifiers (conceptID), which link together instances of the same concept, e.g. all the sentences pertaining to the same method will be linked together with the same conceptID (e.g. "Met1").

If we combine the layers of annotation so as to

Proceedings of the Workshop on Negation and Speculation in Natural Language Processing, pages 1–4,
Uppsala, July 2010.

Table 1: The CoreSC Annotation scheme

Category	Description
Hypothesis	A statement not yet confirmed rather than a factual statement
Motivation	The reasons behind an investigation
Background	Generally accepted background knowledge and previous work
Goal	A target state of the investigation where intended discoveries are made
Object-New	An entity which is a product or main theme of the investigation
Object-New-Advantage	Advantage of an object
Object-New-Disadvantage	Disadvantage of an object
Method-New	Means by which authors seek to achieve a goal of the investigation
Method-New-Advantage	Advantage of a Method
Method-New-Disadvantage	Disadvantage of a Method
Method-Old	A method mentioned pertaining to previous work
Method-Old-Advantage	Advantage of a Method
Method-Old-Disadvantage	Disadvantage of a Method
Experiment	An experimental method
Model	A statement about a theoretical model or framework
Observation	the data/phenomena recorded in an investigation
Result	factual statements about the outputs of an investigation
Conclusion	statements inferred from observations & results relating to research hypothesis

give flat labels, we cater for the categories in table 1.

The CoreSC scheme was accompanied by a set of 45 page guidelines which contain a decision tree, detailed description of the semantics of the categories, 6 rules for pairwise distinction and examples from chemistry papers. These guidelines are available from http://ie-repository.jisc.ac.uk/88/.

3 The CoreSC corpus

We used the CoreSC annotation scheme and the semantic annotation tool SAPIENT [Liakata et al. (2009)] to construct a corpus of 265 annotated papers [Liakata and Soldatova (2009)] from physical chemistry and biochemistry. The CoreSC corpus was developed in two different phases. During phase I, fifteen Chemistry experts were split into five groups of three, each of which annotated eight different papers; A 16th expert annotated across groups as a consistency check. This resulted in a total of 41 papers being annotated, all of which received multiple annotations. We ranked annotators according to median success in terms of inter-annotator agreement (as measured by Cohen's kappa) both within their groups and for a paper common across groups. In phase II, the 9 best annotators of phase I each annotated 25 papers, amounting to a total of 225 papers.

The CoreSC corpus is now being used to train a classifier for the automation of Core Scientific concepts in papers.

4 Correlating CoreSCs to other zones of interest

Given the plethora of annotation schemes, it is interesting to investigate the correlation between different views of scientific papers and how different schemes map to each other. We recently looked at the correlation between the CoreSC scheme, which views papers as the humanly readable representation of scientific investigations and seeks to recover the investigation components within the paper, and AZ-II [Teufel et al. (2009)], which assumes a paper is the attempt of claiming ownership for a new piece of knowledge and aims to recover the rhetorical structure and the relevant stages in the argumentation.

By definition, the two schemes focus on different aspects of the papers, with CoreSCs providing more detail with respect to different types of methods and results and AZ-II looking mostly at the appropriation of knowledge claims. Based on a set of 36 papers annotated with both schemes, we were able to confirm that the two schemes are indeed complementary [Liakata et al. (2010)]. CoreSC categories provide a greater level of granularity when it comes to the content-related categories whereas AZ-II categories cover aspects of the knowledge claims that permeate across different CoreSC concepts.

In [Guo et al. (2010)] we followed a similar methodology for annotating abstracts with CoreSCs and an independently produced annotation scheme for abstract sections [Hirohata et al. (2008)]. We found a subsumption relation between the schemes, with CoreSCs providing the

finer granularity.

To obtain the mapping between annotation schemes, which allows annotation schemes to be defined in a wider context, we ideally require annotations from different schemes to be made available for the same set of papers. However, a first interpretation of the relation between schemes can be made by mapping between annotation guidelines.

5 Thoughts on using CoreSCs for Negation and Speculation

Current work of ours involves automating the recognition of CoreSCs and we plan to use them to produce extractive summaries for papers. We are also in the process of evaluating the usefulness of CoreSCs for Cancer Risk Assessment (CRA). An important aspect of the latter is being able to distinguish between positive and negative results and assess the confidence in any conclusions drawn. This naturally leads us to the need for exploring negation and speculation, both of which are prominent in scientific papers, as well as how these two phenomena correlate to CoreSCs.

While it seems that negation can be identified by means of certain linguistic patterns [Morante (2010)], different types of negation can appear throughout the paper, some pertaining to background work, problems serving as the motivation of the paper, others referring to intermediate results or conclusions. It is interesting to look at these different types of negation in the context of each of the different CoreSCs, the type of linguistic patterns used to express it and their distribution across CoreSCs. This can provide a more targetted approach to negation, while at the same time it can be used in combination with a CoreSC to infer the type of knowledge obtained (e.g. a positive or negative result). We plan to use automatic methods for recognising negation patterns in CoreSCs and relate them to specific CoreSC categories.

There is a consensus that identifying speculation is a harder task than identifying negation. Part of the problem is that "speculative assertions are to be identified on the basis of the judgements about the author's intended meaning, rather than on the presence of certain designated hedge terms" [Medlock and Briscoe (2007); Light et al. (2004)]. When annotating papers with CoreSCs, annotators are required to understand the paper content rather than base category assignments entirely on linguistic patterns. This is why we have chosen experts as annotators for the creation of the CoreSC corpus. So both speculation and CoreSC annotation appear to be higher level annotation tasks requiring comprehension of the intended meaning. Looking at the annotation guidelines for hedges [Medlock and Briscoe (2007)], it would seem that cases of hedge type 1 correspond to to CoreSC `Conclusion`, hedge type 2 pertains to `Background`, hedge type 3 would mainly be cases of `Motivation`, hedge type 4 maps to `Motivation` or `Hypothesis`, hedge type 5 maps to `Goal` and hedge type 6 maps to `Conclusion`. One can look at speculation in the zones/windows identified by the previously mentioned CoreSCs. Indeed, two of the categories, `Hypothesis` and `Motivation` are speculative by definition. We intend to port the issue of identifying speculation in our papers to that of identifying the corresponding CoreSCs. We also plan to annotate the hedge classification data of [Medlock and Briscoe (2007)] with CoreSCs to confirm the mapping between the two schemes.

References

Y. Guo, A. Korhonen, M. Liakata, I Silins, L. LiSun, and U. Stenius. Identifying the information structure of scientific abstracts: An investigation of three different schemes. In *Proceedings of BioNLP 2010. To appear.*, Uppsala, Sweden, 2010.

K. Hirohata, N. Okazaki, S. Ananiadou, and M. Ishizuka. Identifying sections in scientific abstracts using conditional random fields. In *Proc. of the IJCNLP 2008*, 2008.

M. Liakata and L.N. Soldatova. The art corpus. Technical report, Aberystwyth University, 2009. URL `http://www.aber.ac.uk/en/cs/research/cb/projects/art/art-corpus/`.

M. Liakata, Claire Q, and S. Soldatova. Semantic annotation of papers: Interface & enrichment tool (sapient). In *Proceedings of BioNLP-09*, pages 193–200, Boulder, Colorado, 2009.

M. Liakata, S. Teufel, A. Siddharthan, and C. Batchelor. Corpora for the conceptualisation and zoning of scientific papers. 2010.

M. Light, X.Y. Qiu, and P. Srinivasan. The language of bioscience: Facts, speculations and statements in between. In *Proceedings of BioLink 2004 Workshop on Linking Biological Literature, Ontologies and Databases: Tools for Users*, Boston, 2004.

J. Lin. Is searching full text more effective than searching abstracts? *BMC Bioinformatics*, 10:46, 2009.

T. McIntosh and J.R. Curran. Challenges for automatically extracting molecular interactions from full-text articles. *BMC Bioinformatics*, 10(311), 2009.

B. Medlock and T. Briscoe. Weakly supervised learning for hedge classification in scientific literature. In *45th Annual Meeting of the ACL*, pages 23–30, Prague, Czech Republic, 2007.

R. Morante. Descriptive analysis of negation cues in biomedical texts. In *Proceedings of the Seventh International Language Resources and Evaluation (LREC'10)*, pages 1429–1436, Valletta, Malta, 2010.

S. Pyysalo, F. Ginter, J. Heimonen, J. Bjorne, J. Boberg, J. Jarvinen, and T. Salakoski. Bioinfer: a corpus for information extraction in the biomedical domain. *BMC Bioinformatics*, 8(1), 2007.

H. Shatkay, F. Pan, A. Rzhetsky, and W.J. Wilbur. Multi-dimensional classification of biomedical text: Toward automated, practical provision of high-utility text to diverse users. *Journal of Bioinformatics*, 24:18:2086–2093, 2008.

L.N. Soldatova and R.D. King. An ontology of scientific experiments. *Journal of the Royal Society Interface*, 3:795–803, 2006.

L.N. Soldatova and M. Liakata. An ontology methodology and cisp-the proposed core information about scientific papers. Technical Report JISC Project Report, Aberystwyth University, 2007. URL `http://ie-repository.jisc.ac.uk/137/`.

S. Teufel and A. Siddharthan. Whose idea was this, and why does it matter? attributing sicentific work to citations. In *Proceedings of NAACL-HLT-07*, 2007.

Simone Teufel, Advaith Siddharthan, and Colin Batchelor. Towards discipline-independent argumentative zoning: Evidence from chemistry and computational linguistics. In *Proceedings of EMNLP-09*, Singapore, 2009.

V. Vincze, G. Szarvas, R. Farkas, G. Mra, and J. Csirik. The bioscope corpus: biomedical texts annotated for uncertainty, negation and their scopes. *BMC Bioinformatics*, 9(Suppl 11):S9, 2008.

Creating and Evaluating a Consensus for Negated and Speculative Words in a Swedish Clinical Corpus

Hercules Dalianis, Maria Skeppstedt
Department of Computer and Systems Sciences (DSV)
Stockholm University
Forum 100
SE-164 40 Kista, Sweden
{hercules, mariask}@dsv.su.se

Abstract

In this paper we describe the creation of a consensus corpus that was obtained through combining three individual annotations of the same clinical corpus in Swedish. We used a few basic rules that were executed automatically to create the consensus. The corpus contains negation words, speculative words, uncertain expressions and certain expressions. We evaluated the consensus using it for negation and speculation cue detection. We used Stanford NER, which is based on the machine learning algorithm Conditional Random Fields for the training and detection. For comparison we also used the clinical part of the BioScope Corpus and trained it with Stanford NER. For our clinical consensus corpus in Swedish we obtained a precision of 87.9 percent and a recall of 91.7 percent for negation cues, and for English with the Bioscope Corpus we obtained a precision of 97.6 percent and a recall of 96.7 percent for negation cues.

1 Introduction

How we use language to express our thoughts, and how we interpret the language of others, varies between different speakers of a language. This is true for various aspects of a language, and also for the topic of this article; negations and speculations. The differences in interpretation are of course most relevant when a text is used for communication, but it also applies to the task of annotation. When the same text is annotated by more than one annotator, given that the annotating task is non-trivial, the resulting annotated texts will not be identical. This will be the result of differences in how the text is interpreted, but also of differences in how the instructions for annotation are interpreted. In order to use the annotated texts, it must first be decided if the interpretations by the different annotators are similar enough for the purpose of the text, and if so, it must be decided how to handle the non-identical annotations.

In the study described in this article, we have used a Swedish clinical corpus that was annotated for certainty and uncertainty, as well as for negation and speculation cues by three Swedish-speaking annotators. The article describes an evaluation of a consensus annotation obtained through a few basic rules for combining the three different annotations into one annotated text.[1]

2 Related research

2.1 Previous studies on detection of negation and speculation in clinical text

Clinical text often contains reasoning, and thereby many uncertain or negated expressions. When, for example, searching for patients with a specific symptom in a clinical text, it is thus important to be able to detect if a statement about this symptom is negated, certain or uncertain.

The first approach to identifying negations in Swedish clinical text was carried out by Skeppstedt (2010), by whom the well-known NegEx algorithm (Chapman et al., 2001), created for English clinical text, was adapted to Swedish clinical text. Skeppstedt obtained a precision of 70 percent and a recall of 81 percent in identifying negated diseases and symptoms in Swedish clinical text. The NegEx algorithm is purely rule-based, using lists of cue words indicating that a preceding or following disease or symptom is negated. The English version of NegEx (Chapman et al., 2001) obtained a precision of 84.5 percent and a recall of 82.0 percent.

[1] This research has been carried out after approval from the Regional Ethical Review Board in Stockholm (Etikprvningsnmnden i Stockholm), permission number 2009/1742-31/5.

Proceedings of the Workshop on Negation and Speculation in Natural Language Processing, pages 5–13,
Uppsala, July 2010.

Another example of negation detection in English is the approach used by Huang and Lowe (2007). They used both parse trees and regular expressions for detecting negated expressions in radiology reports. Their approach could detect negated expressions both close to, and also at some distance from, the actual negation cue (or what they call negation signal). They obtained a precision of 98.6 percent and a recall of 92.6 percent.

Elkin et al. (2005) used the terms in SNOMED-CT (Systematized Nomenclature of Medicine-Clinical Terms), (SNOMED-CT, 2010) and matched them to 14 792 concepts in 41 health records. Of these concepts, 1 823 were identified as negated by humans. The authors used Mayo Vocabulary Server Parsing Engine and lists of cue words triggering negation as well as words indicating the scope of these negation cues. This approach gave a precision of 91.2 percent and a recall of 97.2 percent in detecting negated SNOMED-CT concepts.

In Rokach et al. (2008), they used clinical narrative reports containing 1 766 instances annotated for negation. The authors tried several machine learning algorithms for detecting negated findings and diseases, including hidden markov models, conditional random fields and decision trees. The best results were obtained with cascaded decision trees, with nodes consisting of regular expressions for negation patterns. The regular expressions were automatically learnt, using the LCS (longest common subsequence) algorithm on the training data. The cascaded decision trees, built with LCS, gave a precision of 94.4 percent, a recall of 97.4 percent and an F-score of 95.9 percent.

Szarvas (2008) describes a trial to automatically identify speculative sentences in radiology reports, using Maximum Entropy Models. Advanced feature selection mechanisms were used to automatically extract cue words for speculation from an initial seed set of cues. This, combined with manual selection of the best extracted candidates for cue words, as well as with outer dictionaries of cue words, yielded an F-score of 82.1 percent for detecting speculations in radiology reports. An evaluation was also made on scientific texts, and it could be concluded that cue words for detecting speculation were domain-specific.

Morante and Daelemans (2009) describe a machine learning system detecting the scope of negations, which is based on meta-learning and is trained and tested on the annotated BioScope Corpus. In the clinical part of the corpus, the authors obtained a precision of 100 percent, a recall of 97.5 percent and finally an F-score of 98.8 percent on detection of cue words for negation. The authors used TiMBL (Tilburg Memory Based Learner), which based its decision on features such as the words annotated as negation cues and the two words surrounding them, as well as the part of speech and word forms of these words. For detection of the negation scope, the task was to decide whether a word in a sentence containing a negation cue was either the word starting or ending a negation scope, or neither of these two. Three different classifiers were used: support vector machines, conditional random fields and TiMBL. Features that were used included the word and the two words preceding and following it, the part of speech of these words and the distance to the negation cue. A fourth classifier, also based on conditional random fields, used the output of the other three classifiers, among other features, for the final decision. The result was a precision of 86.3 percent and a recall of 82.1 percent for clinical text. It could also be concluded that the system was portable to other domains, but with a lower result.

2.2 The BioScope Corpus

Annotated clinical corpora in English for negation and speculation are described in Vincze et al. (2008), where clinical radiology reports (a subset of the so called BioScope Corpus) encompassing 6 383 sentences were annotated for negation, speculation and scope. Henceforth, when referring to the BioScope Corpus, we only refer to the clinical subset of the BioScope Corpus. The authors found 877 negation cues and 1 189 speculation cues, (or what we call speculative cues) in the corpora in 1 561 sentences. This means that fully 24 percent of the sentences contained some annotation for negation or uncertainty. However, of the original 6 383 sentences, 14 percent contained negations and 13 percent contained speculations. Hence some sentences contained both negations and speculations. The corpus was annotated by two students and their work was led by a chief annotator. The students were not allowed to discuss their annotations with each other, except at regular meetings, but they were allowed to discuss

with the chief annotator. In the cases where the two student annotators agreed on the annotation, that annotation was chosen for the final corpus. In the cases where they did not agree, an annotation made by the chief annotator was chosen.

2.3 The Stanford NER based on CRF

The Stanford Named Entity Recognizer (NER) is based on the machine learning algorithm Conditional Random Fields (Finkel et al., 2005) and has been used extensively for identifying named entities in news text. For example in the CoNLL-2003, where the topic was language-independent named entity recognition, Stanford NER CRF was used both on English and German news text for training and evaluation. Where the best results for English with Stanford NER CRF gave a precision of 86.1 percent, a recall of 86.5 percent and F-score of 86.3 percent, for German the best results had a precision of 80.4 percent, a recall of 65.0 percent and an F-score of 71.9 percent, (Klein et al., 2003). We have used the Stanford NER CRF for training and evaluation of our consensus.

2.4 The annotated Swedish clinical corpus for negation and speculation

A process to create an annotated clinical corpus for negation and speculation is described in Dalianis and Velupillai (2010). A total of 6 740 randomly extracted sentences from a very large clinical corpus in Swedish were annotated by three non-clinical annotators. The sentences were extracted from the text field Assessment (*Bedömning* in Swedish). Each sentence and its context from the text field Assessment were presented to the annotators who could use five different annotation classes to annotate the corpora. The annotators had discussions every two days on the previous days' work led by the experiment leader.

As described in Velupillai (2010), the annotation guidelines were inspired by the BioScope Corpus guidelines. There were, however, some differences, such as the scope of a negation or of an uncertainty not being annotated. It was instead annotated if a sentence or clause was certain, uncertain or undefined. The annotators could thus choose to annotate the entire sentence as belonging to one of these three classes, or to break up the sentence into subclauses.

Pairwise inter-annotator agreement was also measured in the article by Dalianis and Velupillai (2010) . The average inter-annotator agreement increased after the first annotation rounds, but it was lower than the agreement between the annotators of the BioScope Corpus.

The annotation classes used were thus *negation* and *speculative words*, but also *certain expression* and *uncertain expression* as well as *undefined*. The annotated subset contains a total of 6 740 sentences or 71 454 tokens, including its context.

3 Method for constructing the consensus

We constructed a consensus annotation out of the three different annotations of the same clinical corpus that is described in Dalianis and Velupillai (2010). The consensus was constructed with the general idea of choosing, as far as possible, an annotation for which there existed an identical annotation performed by at least two of the annotators, and thus to find a majority annotation. In the cases where no majority was found, other methods were used.

Other options would be to let the annotators discuss the sentences that were not identically annotated, or to use the method of the BioScope Corpus, where the sentences that were not identically annotated were resolved by a chief annotator (Vincze et al., 2008). A third solution, which might, however, lead to a very biased corpus, would be to not include the sentences for which there was not a unanimous annotation in the resulting consensus corpus.

3.1 The creation of a consensus

The annotation classes that were used for annotation can be divided into two levels. The first level consisted of the annotation classes for classifying the type of sentence or clause. This level thus included the annotation classes *uncertain*, *certain* and *undefined*. The second level consisted of the annotation classes for annotating cue words for negation and speculation, thus the annotation classes *negation* and *speculative words*. The annotation classes on the first level were considered as more important for the consensus, since if there was no agreement on the kind of expression, it could perhaps be said to be less important which cue phrases these expressions contained. In the following constructed example, the annotation tag *Uncertain* is thus an annotation on the first level, while the annotation tags *Negation* and *Speculative words* are on the second level.

```
<Sentence>
    <Uncertain>
      <Speculative_words>
        <Negation>Not</Negation>
        really
      </Speculative_words>
      much worse than before
    </Uncertain>
 <Sentence>
```

When constructing the consensus corpus, the annotated sentences from the first rounds of annotation were considered as sentences annotated before the annotators had fully learnt to apply the guidelines. The first 1 099 of the annotated sentences, which also had a lower inter-annotator agreement, were therefore not included when constructing the consensus. Thereby, 5 641 sentences were left to compare.

The annotations were compared on a sentence level, where the three versions of each sentence were compared. First, sentences for which there existed an identical annotation performed by at least two of the annotators were chosen. This was the case for 5 097 sentences, thus 90 percent of the sentences.

For the remaining 544 sentences, only annotation classes on the first level were compared for a majority. For the 345 sentences where a majority was found on the first level, a majority on the second level was found for 298 sentences when the scope of these tags was disregarded. The annotation with the longest scope was then chosen. For the remaining 47 sentences, the annotation with the largest number of annotated instances on the second level was chosen.

The 199 sentences that were still not resolved were then once again compared on the first level, this time disregarding the scope. Thereby, 77 sentences were resolved. The annotation with the longest scopes on the first-level annotations was chosen.

The remaining 122 sentences were removed from the consensus. Thus, of the 5 641 sentences, 2 percent could not be resolved with these basic rules. In the resulting corpus, 92 percent of the sentences were identically annotated by at least two persons.

3.2 Differences between the consensus and the individual annotations

Aspects of how the consensus annotation differed from the individual annotations were measured. The number of occurrences of each annotation class was counted, and thereafter normalised on the number of sentences, since the consensus annotation contained fewer sentences than the original, individual annotations.

The results in Table 1 show that there are fewer uncertain expressions in the consensus annotation than in the average of the individual annotations. The reason for this could be that if the annotation is not completeley free of randomness, the class with a higher probability will be more frequent in a majority consensus, than in the individual annotations. In the cases where the annotators are unsure of how to classify a sentence, it is not unlikely that the sentence has a higher probability of being classified as belonging to the majority class, that is, the class *certain*.

The class *undefined* is also less common in the consensus annotation, and the same reasoning holds true for *undefined* as for *uncertain*, perhaps to an even greater extent, since *undefined* is even less common.

Also the *speculative* words are fewer in the consensus. Most likely, this follows from the *uncertain* sentences being less common.

The words annotated as *negations*, on the other hand, are more common in the consensus annotation than in the individual annotations. This could be partly explained by the choice of the 47 sentences with an annotation that contained the largest number of annotated instances on the second level, and it is an indication that the consensus contains some annotations for negation cues which have only been annotated by one person.

Type of Annot. class	**Individ.**	**Consens.**
Negation	853	910
Speculative words	1 174	1 077
Uncertain expression	697	582
Certain expression	4 787	4 938
Undefined expression	257	146

Table 1: Comparison of the number of occurrences of each annotation class for the individual annotations and the consensus annotation. The figures for the individual annotations are the mean of the three annotators, normalised on the number of sentences in the consensus.

Table 2 shows how often the annotators have divided the sentences into clauses and annotated each clause with a separate annotation class. From the table we can see that annotator A and also an-

notator H broke up sentences into more than one type of the expressions *Certain*, *Uncertain* or *Undefined expressions* more often than annotator F. Thereby, the resulting consensus annotation has a lower frequency of sentences that contained these annotations than the average of the individual annotations. Many of the more granular annotations that break up sentences into certain and uncertain clauses are thus not included in the consensus annotation. There are instead more annotations that classify the entire sentence as either *Certain*, *Uncertain* or *Undefined*.

Annotators	**A**	**F**	**H**	**Cons.**
No. sentences	349	70	224	147

Table 2: Number of sentences that contained more than one instance of either one of the annotation classes *Certain*, *Uncertain* or *Undefined expressions* or a combination of these three annotation classes.

3.3 Discussion of the method

The constructed consensus annotation is thus different from the individual annotations, and it could at least in some sense be said to be better, since 92 percent of the sentences have been identically annotated by at least two persons. However, since for example some expressions of uncertainty, which do not have to be incorrect, have been removed, it can also be said that some information containing possible interpretations of the text, has also been lost.

The applied heuristics are in most cases specific to this annotated corpus. The method is, however, described in order to exemplify the more general idea to use a majority decision for selecting the correct annotations. What is tested when using the majority method described in this article for deciding which annotation is correct, is the idea that a possible alternative to a high annotator agreement would be to ask many annotators to judge what they consider to be certain or uncertain. This could perhaps be based on a very simplified idea of language, that the use and interpretation of language is nothing more than a majority decision by the speakers of that language.

A similar approach is used in Steidl et al. (2005), where they study emotion in speech. Since there are no objective criteria for deciding with what emotion something is said, they use manual classification by five labelers, and a majority voting for deciding which emotion label to use. If less than three labelers agreed on the classification, it was omitted from the corpus.

It could be argued that this is also true for uncertainty, that if there is no possibility to ask the author of the text, there are no objective criteria for deciding the level of certainty in the text. It is always dependent on how it is perceived by the reader, and therefore a majority method is suitable. Even if the majority approach can be used for subjective classifications, it has some problems. For example, to increase validity more annotators are needed, which complicates the process of annotation. Also, the same phenomenon that was observed when constructing the consensus would probably also arise, that a very infrequent class such as *uncertain*, would be less frequent in the majority consensus than in the individual annotations. Finally, there would probably be many cases where there is no clear majority for either completely certain or uncertain: in these cases, having many annotators will not help to reach a decision and it can only be concluded that it is difficult to classify this part of a text. Different levels of uncertainty could then be introduced, where the absence of a clear majority could be an indication of weak certainty or uncertainty, and a very weak majority could result in an undefined classification.

However, even though different levels of certainty or uncertainty are interesting when studying how uncertainties are expressed and perceived, they would complicate the process of information extraction. Thus, if the final aim of the annotation is to create a system that automatically detects what is certain or uncertain, it would of course be more desirable to have an annotation with a higher inter-annotator agreement. One way of achieving a this would be to provide more detailed annotation guidelines for what to define as certainty and uncertainty. However, when it comes to such a vague concept as uncertainty, there is always a thin line between having guidelines capturing the general perception of uncertainty in the language and capturing a definition of uncertainty that is specific to the writers of the guidelines. Also, there might perhaps be a risk that the complex concept of certainty and uncertainty becomes overly simplified when it has to be formulated as a limited set of guidelines. Therefore, a more feasible method of achieving higher agreement is probably to instead

Class Neg-Spec	Relevant	Retrieved	Corpus	Precision	Recall	F-score
Negation	782	890	853	0.879	0.917	0.897
Speculative words	376	558	1061	0.674	0.354	0.464
Total	1 158	1 448	1 914	0.800	0.605	0.687

Table 3: The results for *negation* and *speculation* on consensus when executing Stanford NER CRF using ten-fold cross validation.

Class Cert-Uncertain	Relevant	Retrieved	Corpus	Precision	Recall	F-score
Certain expression	4 022	4 903	4 745	0.820	0.848	0.835
Uncertain expression	214	433	577	0.494	0.371	0.424
Undefined expression	2	5	144	0.400	0.014	0.027
Total	4 238	5 341	5 466	0.793	0.775	0.784

Table 4: The results for *certain* and *uncertain* on consensus when executing Stanford NER CRF using ten-fold cross validation.

simplify what is being annotated, and not annotate for such a broad concept as uncertainty in general.

Among other suggestions for improving the annotation guidelines for the corpus that the consensus is based on, Velupillai (2010) suggests that the guidelines should also include instructions on the focus of the uncertainties, that is, what concepts are to be annotated for uncertainty.

The task could thus, for example, be tailored towards the information that is to be extracted, and thereby be simplified by only annotating for uncertainty relating to a specific concept. If diseases or symptoms that are present in a patient are to be extracted, the most relevant concept to annotate is whether a finding is present or not present in the patient, or whether it is uncertain if it is present or not. This approach has, for example, achieved a very high inter-annotator agreement in the annotation of the evaluation data used by Chapman et al. (2001). Even though this approach is perhaps linguistically less interesting, not giving any information on uncertainties in general, if the aim is to search for diseases and symptoms in patients, it should be sufficient.

In light of the discussion above, the question to what extent the annotations in the constructed consensus capture a general perception of certainty or uncertainty must be posed. Since it is constructed using a majority method with three annotators, who had a relatively low pairwise agreement, the corpus could probably not be said to be a precise capture of what is a certainty or uncertainty. However, as Artstein and Poesio (2008) point out, it cannot be said that there is a fixed level of agreement that is valid for all purposes of a corpus, but the agreement must be high enough for a certain purpose. Therefore, if the information on whether there was a unanimous annotation of a sentence or not is retained, serving as an indicator of how typical an expression of certainty or uncertainty is, the constructed corpus can be a useful resource. Both for studying how uncertainty in clinical text is constructed and perceived, and as one of the resources that is used for learning to automatically detect certainty and uncertainty in clinical text.

4 Results of training with Stanford NER CRF

As a first indication of whether it is possible to use the annotated consensus corpus for finding negation and speculation in clinical text, we trained the Stanford NER CRF, (Finkel et al., 2005) on the annotated data. Artstein and Poesio (2008) write that the fact that annotated data can be generalized and learnt by a machine learning system is not an indication that the annotations capture some kind of reality. If it would be shown that the constructed consensus is easily generalizable, this can thus not be used as an evidence of its quality. However, if it would be shown that the data obtained by the annotations cannot be learnt by a machine learning system, this can be used as an indication that the data is not easily generalizable and that the task to learn perhaps should, if possible, be simplified. Of course, it could also be an indication that another learning algorithm should be used or other features selected.

We created two training sets of annotated consensus material.

The first training set contained annotations on the second level, thus annotations that contained the classes *Speculative words* and *Negation*. In 76 cases, the tag for *Negation* was inside an annotation for *Speculative words*, and these occurrences

Class Neg-Spec Bio	Relevant	Retrieved	Corpus	Precision	Recall	F-score
Negation	843	864	872	0.976	0.967	0.971
Speculative words	1 021	1 079	1 124	0.946	0.908	0.927
Scope[1]	1 295	1 546	1 595[2]	0.838	0.812	0.825

Table 5: The results for *negations, speculation cues* and *scopes* on the BioScope Corpus when executing Stanford NER CRF using ten-fold cross validation.

Class Neg-Spec	Relevant	Retrieved	Corpus	Precision	Recall	F-score
Negation A	791	1 005	896	0.787	0.883	0.832
Speculative words	684	953	1 699	0.718	0.403	0.516
Negation F	938	1097	1023	0.855	0.916	0.884
Speculative words	464	782	1 496	0.593	0.310	0.407
Negation H	722	955	856	0.756	0.843	0.797
Speculative words	552	853	1 639	0.647	0.336	0.443

Table 6: The results for *negations* and *speculation cues* and *scopes* for annotator A, F and H respectively when executing Stanford NER CRF using ten-fold cross validation.

of the tag *Negation* were removed. It is detecting this difference between a real negation cue and a negation word inside a cue for speculation that is one of the difficulties that distinguishes the learning task from a simple string matching.

The second training set only contained the consensus annotations on the first level, thus the annotation classes *Certain*, *Uncertain* and *Undefined*.

We used the default settings on Stanford NER CRF. The results of the evaluation using ten-fold cross validation (Kohavi, 1995) are shown in Table 3 and Table 4.

As a comparison, and to verify the suitability of the chosen machine learning method, we also trained and evaluated the BioScope Corpus using Stanford NER CRF for negation, speculation and scope. The results can be seen in Table 5. When training the detection of scope, only BioScope sentences that contained an annotation for negation and speculation were selected for the training and evaluation material for the Stanford NER CRF. This division into two training sets follows the method used by Morante and Daelemans (2009), where sentences containing a cue are first detected, and then, among these sentences, the scope of the cue is determined.

We also trained and evaluated the annotations that were carried out by each annotator A, F and H separately, i.e. the source of consensus. The results can be seen in Table 6.

We also compared the distribution of *Negation* and *Speculative words* in the consensus versus the BioScope Corpus and we found that the consensus, in Swedish, used about the same number of (types) for negation as the BioScope Corpus in English (see Table 7), but for *speculative words* the consensus contained many more types than the BioScope Corpus. In the constructed consensus, 72 percent of the *Speculative words* occurred only once, whereas in the BioScope Corpus this was the case for only 24 percent of the *Speculative words*.

Type of word	Cons.	Bio
Unique words (Types) annotated as *Negation*	13	19
Negations that occurred only once	5	10
Unique words (Types) annotated as *Speculative*	408	79
Speculative words that occurred only once	294	19

Table 7: Number of unique words both in the Consensus and in the BioScope Corpus that were annotated as *Negation* and as *Speculative words*, and how many of these that occurred only once.

5 Discussion

The training results using our clinical consensus corpus in Swedish gave a precision of 87.9 percent and a recall of 91.7 percent for negation cues and a precision of 67.4 percent and a recall of 35.4 percent for speculation cues. The results for detecting negation cues are thus much higher than for detecting cues for speculation using Stanford NER CRF. This difference is not very surprising, given

[1]The scopes were trained and evaluated separetely from the negations and speculations.

[2]The original number of annotated scopes in the BioScope Corpus is 1 981. Of these, 386 annotations for nested scopes were removed.

the data in Table 7, which shows that there are only a very limited number of negation cues, whereas there exist over 400 different cue words for speculation. One reason why the F-score for negation cues is not even higher, despite the fact that the number of cues for negations is very limited, could be that a negation word inside a tag for *speculative words* is not counted as a negation cue. Therefore, the word *not* in, for example, *not really* could have been classified as a negation cue by Stanford NER CRF, even though it is a cue for speculation and not for negation. Another reason could be that the word meaning *without* in Swedish (*utan*) also means *but*, which only sometimes makes it a negation cue.

We can also observe in Table 4, that the results for detection of uncertain expressions are very low (F-score 42 percent). For undefined expressions, due to scarce training material, it is not possible to interpret the results. For certain expressions the results are acceptable, but since the instances are in majority, the results are not very useful.

Regarding the BioScope Corpus we can observe (see Table 5) that the training results both for detecting cues for negation and for speculations are very high, with an F-score of 97 and 93 percent, respectively. For scope detection, the result is lower but acceptable, with an F-score of 83 percent. These results indicate that the chosen method is suitable for the learning task.

The main reason for the differences in F-score between the Swedish consensus corpus and the BioScope Corpus, when it comes to the detection of speculation cues, is probably that the variation of words that were annotated as *Speculative word* is much larger in the constructed consensus than in the BioScope Corpus.

As can be seen in Table 7, there are many more types of speculative words in the Swedish consensus than in the BioScope Corpus. We believe that one reason for this difference is that the sentences in the constructed consensus are extracted from a very large number of clinics (several hundred), whereas the BioScope Corpus comes from one radiology clinic. This is supported by the findings of Szarvas (2008), who writes that cues for speculation are domain-specific. In this case, however, the texts are still within the domain of clinical texts.

Another reason for the larger variety of cues for speculation in the Swedish corpus could be that the guidelines for annotating the BioScope Corpus and the method for creating a consensus were different.

When comparing the results for the individual annotators with the constructed consensus, the figures in Tables 3 and 6 indicate that there are no big differences in generalizability. When detecting cues for negation, the precision for the consensus is better than the precision for the individual annotations. However, the results for the recall are only slightly better or equivalent for the consensus than for the individual annotations. If we analyse the speculative cues we can observe that the consensus and the individual annotations have similar results.

The low results for learning to detect cues for speculation also serve as an indicator that the task should be simplified to be more easily generalizable. For example, as previously suggested for increasing the inter-annotator agreement, the task could be tailored towards the specific information that is to be extracted, such as the presence of a disease in a patient.

6 Future work

To further investigate if a machine learning algorithm such as Conditional Random Fields can be used for detecting speculative words, more information needs to be provided for the Conditional Random Fields, such as part of speech or if any of the words in the sentence can be classified as a symptom or a disease. One Conditional Random Fields system that can treat nested annotations is CRF++ (CRF++, 2010). CRF++ is used by several research groups and we are interested in trying it out for the negation and speculation detection as well as scope detection.

7 Conclusion

A consensus clinical corpus was constructed by applying a few basic rules for combining three individual annotations into one. Compared to the individual annotations, the consensus contained fewer annotations of uncertainties and fewer annotations that divided the sentences into clauses. It also contained fewer annotations for speculative words, and more annotations for negations. Of the sentences in the constructed corpus, 92 percent were identically annotated by at least two persons.

In comparison with the BioScope Corpus, the constructed consensus contained both a larger number and a larger variety of speculative cues.

This might be one of the reasons why the results for detecting cues for speculative words using the Stanford NER CRF are much better for the BioScope Corpus than for the constructed consensus corpus; the F-scores are 93 percent versus 46 percent.

Both the BioScope Corpus and the constructed consensus corpus had high values for detection of negation cues, F-scores 97 and 90 percent, respectively.

As is suggested by Velupillai (2010), the guidelines for annotation should include instructions on the focus of the uncertainties. To focus the decision of uncertainty on, for instance, the disease of a patient, might improve both the inter-annotator agreement and the possibility of automatically learning to detect the concept of uncertainty.

Acknowledgments

We are very grateful for the valuable comments by the three anonymous reviewers.

References

Ron Artstein and Massimo Poesio. 2008. Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4):555–596.

Wendy W. Chapman, Will Bridewell, Paul Hanbury, Gregory F. Cooper, and Bruce G. Buchanan. 2001. A simple algorithm for identifying negated findings and diseases in discharge summaries. *Journal of biomedical informatics*, 34(5):301–310.

CRF++. 2010. CRF++: Yet another CRF toolkit, May 8. http://crfpp.sourceforge.net//.

Hercules Dalianis and Sumithra Velupillai. 2010. How certain are clinical assessments? Annotating Swedish clinical text for (un)certainties, speculations and negations. In *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*, Valletta, Malta, May.

Peter L. Elkin, Steven H. Brown, Brent A. Bauer, Casey S. Husser, William Carruth, Larry R. Bergstrom, and Dietlind L. Wahner-Roedler. 2005. A controlled trial of automated classification of negation from clinical notes. *BMC Medical Informatics and Decision Making*, 5(1):13.

Jenny Rose Finkel, Trond Grenager, and Christopher Manning. 2005. Incorporating non-local information into information extraction systems by gibbs sampling. In *ACL '05: Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pages 363–370.

Yang Huang and Henry J. Lowe. 2007. A novel hybrid approach to automated negation detection in clinical radiology reports. *Journal of the American Medical Informatics Association*, 14(3):304.

Dan Klein, Joseph Smarr, Huy Nguyen, and Christopher D. Manning. 2003. Named entity recognition with character-level models. In *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003*, pages 180–183. Association for Computational Linguistics.

Ron Kohavi. 1995. A study of cross-validation and bootstrap for accuracy estimation and model selection. In *International Joint Conference on Artificial Intelligence*, volume 14, pages 1137–1145.

Roser Morante and Walter Daelemans. 2009. A metalearning approach to processing the scope of negation. In *CoNLL '09: Proceedings of the Thirteenth Conference on Computational Natural Language Learning*, pages 21–29. Association for Computational Linguistics.

Lior Rokach, Roni Romano, and Oded Maimo. 2008. Negation recognition in medical narrative reports. *Information Retrieval*, 11(6):499–538.

Maria Skeppstedt. 2010. Negation detection in Swedish clinical text. In *Louhi'10 - Second Louhi Workshop on Text and Data Mining of Health Documents, held in conjunction with NAACL HLT 2010*, Los Angeles, June.

SNOMED-CT. 2010. Systematized nomenclature of medicine-clinical terms, May 8. http://www.ihtsdo.org/snomed-ct/.

Stefan Steidl, Michael Levit, Anton Batliner, Elmar Nöth, and Heinrich Niemann. 2005. "Off all things the measure is man" Automatic classification of emotions and inter-labeler consistency. In *Proceeding of the IEEE ICASSP,2005*, pages 317–320.

György Szarvas. 2008. Hedge classification in biomedical texts with a weakly supervised selection of keywords. In *Proceedings of ACL-08: HLT*, pages 281–289, Columbus, Ohio, June. Association for Computational Linguistics.

Sumithra Velupillai. 2010. Towards a better understanding of uncertainties and speculations in swedish clinical text – analysis of an initial annotation trial. To be published in the proceedings of the Negation and Speculation in Natural Language Processing Workshop, July 10, 2010, Uppsala, Sweden.

Veronika Vincze, György Szarvas, Richárd Farkas, György Móra, and János Csirik. 2008. The bioscope corpus: biomedical texts annotated for uncertainty, negation and their scopes. *BMC Bioinformatics*, 9(S-11).

Towards A Better Understanding of Uncertainties and Speculations in Swedish Clinical Text – Analysis of an Initial Annotation Trial

Sumithra Velupillai
Department of Computer and Systems Sciences (DSV)
Stockholm University
Forum 100
SE-164 40 Kista, Sweden
sumithra@dsv.su.se

Abstract

Electronic Health Records (EHRs) contain a large amount of free text documentation which is potentially very useful for Information Retrieval and Text Mining applications. We have, in an initial annotation trial, annotated 6 739 sentences randomly extracted from a corpus of Swedish EHRs for sentence level (un)certainty, and token level speculative keywords and negations. This set is split into different clinical practices and analyzed by means of descriptive statistics and pairwise Inter-Annotator Agreement (IAA) measured by F_1-score. We identify *geriatrics* as a clinical practice with a low average amount of uncertain sentences and a high average IAA, and *neurology* with a high average amount of uncertain sentences. Speculative words are often n-grams, and uncertain sentences longer than average. The results of this analysis is to be used in the creation of a new annotated corpus where we will refine and further develop the initial annotation guidelines and introduce more levels of dimensionality. Once we have finalized our guidelines and refined the annotations we plan to release the corpus for further research, after ensuring that no identifiable information is included.

1 Introduction

Electronic Health Records (EHRs) contain a large amount of free text documentation which is potentially very useful for Information Retrieval and Text Mining applications. Clinical documentation is specific in many ways; there are many authors in a document (e.g. physicians, nurses), there are different situations that are documented (e.g. admission, current status). Moreover, they may often be written under time pressure, resulting in fragmented, brief texts often containing spelling errors and abbreviations. With access to EHR data, many possibilities to exploit documented clinical knowledge and experience arise.

One of the properties of EHRs is that they contain reasoning about the status and diagnoses of patients. Gathering such information for the use in e.g. medical research in order to find relationships between diagnoses, treatments etc. has great potential. However, in many situations, clinicians might describe uncertain or negated findings, which is crucial to distinguish from positive or asserted findings. Potential future applications include search engines where medical researchers can search for particular diseases where negated or speculative contexts are separated from asserted contexts, or text mining systems where e.g. diseases that seem to occur often in speculative contexts are presented to the user, indicating that more research is needed. Moreover, laymen may also benefit from information retrieval systems that distinguish diseases or symptoms that are more or less certain given current medical expertise and knowledge.

We have, in an initial annotation trial, annotated 6 739 sentences randomly extracted from a corpus of Swedish EHRs for sentence level (un)certainty, and token level speculative keywords and negations[1]. In this paper, a deeper analysis of the resulting annotations is performed. The aims are to analyze the results *split into different clinical practices* by means of descriptive statistics and pairwise Inter-Annotator Agreement (IAA) measured by F_1-score, with the goal of identifying a) whether specific clinical practices contain higher or lower amounts of uncertain expressions, b)

[1]This research has been carried out after approval from the Regional Ethical Review Board in Stockholm (Etikprövningsnämnden i Stockholm), permission number 2009/1742-31/5

Proceedings of the Workshop on Negation and Speculation in Natural Language Processing, pages 14–22,
Uppsala, July 2010.

whether specific clinical practices result in higher or lower IAA - indicating a less or more difficult clinical practice for judging uncertainties, and c) identifying the characteristics of the entities annotated as speculative words, are they highly lexical or is a deeper syntactic and/or semantic analysis required for modeling? From this analysis, we plan to conduct a new annotation trial where we will refine and further develop the annotation guidelines and use domain experts for annotations in order to be able to create a useful annotated corpus modeling uncertainties, negations and speculations in Swedish clinical text, which can be used to develop tools for the automatic identification of these phenomena in, for instance, Text Mining applications.

2 Related Research

In recent years, the interest for identifying and modeling speculative language in natural language text has grown. In particular, biomedical scientific articles and abstracts have been the object of several experiments. In Light et al. (2004), four annotators annotated 891 sentences each as either highly speculative, low speculative, or definite, in biomedical scientific abstracts extracted from Medline. In total, they found 11 percent speculative sentences, resulting in IAA results, measured with kappa, between 0.54 and 0.68. One of their main findings was that the majority of the speculative sentences appeared towards the end of the abstract.

Vincze et al. (2008) describe the creation of the BioScope corpus, where more than 20 000 sentences from both medical (clinical) free texts (radiology reports), biological full papers and biological scientific abstracts have been annotated with speculative and negation keywords along with their scope. Over 10 percent of the sentences were either speculative or negated. In the clinical sub-corpus, 14 percent contained speculative keywords. Three annotators annotated the corpus, and the guidelines were modified several times during the annotation process, in order to resolve problematic issues and refine definitions. The IAA results, measured with F_1-score, in the clinical sub-corpus for negation keywords ranged between 0.91 and 0.96, and for speculative keywords between 0.84 and 0.92. The BioScope corpus has been used to train and evaluate automatic classifiers (e.g. Özgür and Radev (2009) and Morante and Daelemans (2009)) with promising results.

Five qualitative dimensions for characterizing scientific sentences are defined in Wilbur et al. (2006), including levels of certainty. Here, guidelines are also developed over a long period of time (more than a year), testing and revising the guidelines consecutively. Their final IAA results, measured with F_1-score, range between 0.70 and 0.80. Different levels of dimensionality for categorizing certainty (in newspaper articles) is also presented in Rubin et al. (2006).

Expressions for communicating probabilities or levels of certainty in clinical care may be inherently difficult to judge. Eleven observers were asked to indicate the level of probability of a disease implied by eighteen expressions in the work presented by Hobby et al. (2000). They found that expressions indicating intermediate probabilities were much less consistently rated than those indicating very high or low probabilities. Similarly, Khorasani et al. (2003) performed a survey analyzing agreement between radiologists and non-radiologists regarding phrases used to convey degrees of certainty. In this study, they found little or no agreement among the survey participants regarding the diagnostic certainty associated with these phrases. Although we do not have access to radiology reports in our corpus, these findings indicate that it is not trivial to classify uncertain language in clinical documentation, even for domain experts.

3 Method

The annotation trial is based on sentences randomly extracted from a corpus of Swedish EHRs (see Dalianis and Velupillai (2010) for an initial description and analysis). These records contain both structured (e.g. measure values, gender information) and unstructured information (i.e. free text). Each free text entry is written under a specific heading, e.g. *Status*, *Current medication*, *Social Background*. For this corpus, sentences were extracted only from the free text entry *Assessment* (Bedömning), with the assumption that these entries contain a substantial amount of reasoning regarding a patient's diagnosis and situation. A simple sentence tokenizing strategy was employed, based on heuristic regular expressions[2]. We have used Knowtator (Ogren, 2006) for the annotation

[2]The performance of the sentence tokenizer has not been evaluated in this work.

work.

One senior level student (SLS), one undergraduate computer scientist (UCS), and one undergraduate language consultant (ULC) annotated the sentences into the following classes; on a sentence level: *certain*, *uncertain* or *undefined*, and on a token level: *speculative words*, *negations*, and *undefined words*.

The annotators are to be considered *naive* coders, as they had no prior knowledge of the task, nor any clinical background. The annotation guidelines were inspired by those created for the BioScope corpus (Vincze et al., 2008), with some modifications (see Dalianis and Velupillai (2010)). The annotators were allowed to break a sentence into subclauses if they found that a sentence contained conflicting levels of certainty, and they were allowed to mark question marks as speculative words. They did not annotate the linguistic scopes of each token level instance. The annotators worked independently, and met for discussions in even intervals (in total seven), in order to resolve problematic issues. No information about the clinic, patient gender, etc. was shown. The annotation trial is considered as a first step in further work of annotating Swedish clinical text for speculative language.

Clinical practice	# sentences	# tokens
hematology	140	1 494
surgery	295	3 269
neurology	351	4 098
geriatrics	142	1 568
orthopaedics	245	2 541
rheumatology	384	3 348
urology	120	1 393
cardiology	128	1 242
oncology	550	5 262
ENT	224	2 120
infection	107	1 228
emergency	717	6 755
paediatrics	935	8 926
total, clinical practice	4 338	43 244
total, full corpus	6 739	69 495

Table 1: Number of sentences and tokens per clinical practice (#sentences > 100), and in total. ENT = Ear, Nose and Throat.

3.1 Annotations and clinical practices

The resulting corpus consists of 6 739 sentences, extracted from 485 unique clinics. In order to be able to analyze possible similarities and differences across clinical practices, sentences from clinics belonging to a specific practice type were grouped together. In Table 1, the resulting groups, along with the total amount of sentences and tokens, are presented[3]. Only groups with a total amount of sentences > 100 were used in the analysis, resulting in 13 groups. A clinic was included in a clinical practice group based on a priority heuristics, e.g. the clinic "Barnakuten-kir" (*Paediatric emergency surgery*) was grouped into paediatrics.

The average length (in tokens) per clinical practice and in total are given in Table 2. Clinical documentation is often very brief and fragmented, for most clinical practices (except urology and cardiology) the minimum sentence length (in tokens) was one, e.g. "basal", "terapisvikt" (*therapy failure*), "lymfödem" (*lymphedema*), "viros" (*virosis*), "opanmäles" (*reported to surgery*, compound with abbreviation). We see that the average sentence length is around ten for all practices, where the shortest are found in rheumatology and the longest in infection.

As the annotators were allowed to break up sentences into subclauses, but not required to, this led to a considerable difference in the total amount of annotations per annotator. In order to be able to analyze similarities and differences between the resulting annotations, all sentence level annotations were converted into *one* sentence class only, the primary class (defined as the first sentence level annotation class, i.e. if a sentence was broken into two clauses by an annotator, the first being *certain* and the second being *uncertain*, the final sentence level annotation class will be *certain*). The sentence level annotation class *certain* was in clear majority among all three annotators. On both sentence and token level, the class *undefined* (a sentence that could not be classified as *certain* or *uncertain*, or a token which was not clearly speculative) was rarely used. Therefore, all sentence level annotations marked as *undefined* are converted to the majority class, *certain*, resulting in two sentence level annotation classes (*certain* and *uncertain*) and two token level annotation classes (*speculative words* and *negations*, i.e. to-

[3] White space tokenization.

kens annotated as *undefined* are ignored).

For the remaining analysis, we focus on the distributions of the annotation classes *uncertain* and *speculative words*, per annotator and annotator pair, and per clinical practice.

Clinical practice	Max	Avg	Stddev
hematology	40	10.67	7.97
surgery	57	11.08	8.29
neurology	105	11.67	10.30
geriatrics	58	11.04	9.29
orthopaedics	40	10.37	6.88
rheumatology	59	8.72	7.99
urology	46	11.61	7.86
cardiology	50	9.70	7.46
oncology	54	9.57	7.75
ENT	54	9.46	7.53
infection	37	11.48	7.76
emergency	55	9.42	6.88
paediatrics	68	9.55	7.24
total, full corpus	120	10.31	8.53

Table 2: Token statistics per sentence and clinical practice. All clinic groups except urology (min = 2) and cardiology (min = 2) have a minimum sentence length of one token.

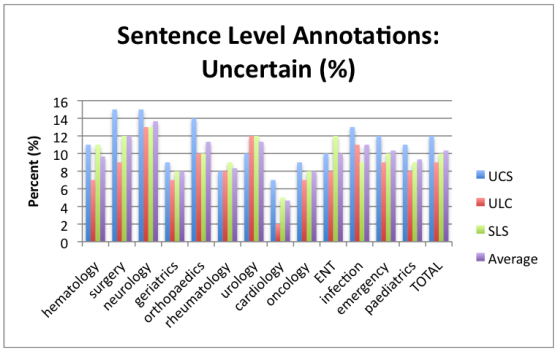

Figure 1: Sentence level annotation: *uncertain*, percentage per annotator and clinical practice.

4 Results

We have measured the proportions (in percent) per annotator for each clinical practice and in total. This enables an analysis of whether there are substantial individual differences in the distributions, indicating that this annotation task is highly subjective and/or difficult. Moreover, we measure IAA by pairwise F_1-score. From this, we may draw conclusions whether specific clinical practices are harder or easier to judge *reliably* (i.e. by high IAA results).

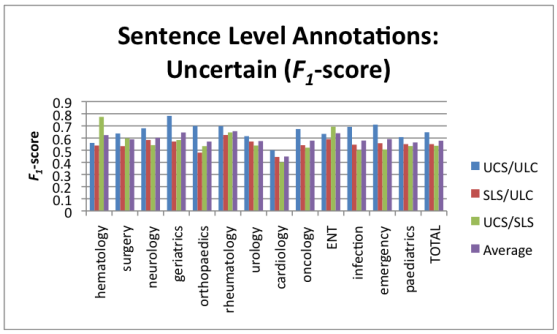

Figure 2: Pairwise F_1-score, sentence level annotation class *uncertain*.

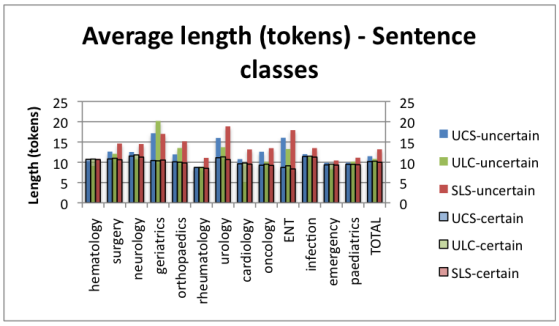

Figure 3: Average length in tokens, per annotator and sentence class.

In Figure 1, we see that the average amount of uncertain sentences lies between 9 and 12 percent for each annotator in the full corpus. In general, UCS has annotated a larger proportion of uncertain sentences compared to ULC and SLS.

The clinical discipline with the highest average amount of uncertain sentences is *neurology* (13.7 percent), the lowest average amount is found in *cardiology* (4.7 percent). Surgery and cardiology show the largest individual differences in proportions (from 9 percent (ULC) to 15 percent (UCS), and from 2 percent (ULC) to 7 percent (UCS), respectively).

However, in Figure 2, we see that the pairwise IAA, measured by F_1-score, is relatively low, with an average IAA of 0.58, ranging between 0.54 (UCS/SLS) and 0.65 (UCS/ULC), for the entire corpus. In general, the annotator pair UCS/ULC have higher IAA results, with the highest for *geriatrics* (0.78). The individual proportions for un-

certain sentences in *geriatrics* is also lower for all annotators (see Figure 1), indicating a clinical practice with a low amount of uncertain sentences, and a slightly higher average IAA (0.64 F_1-score).

4.1 Sentence lengths

As the focus lies on analyzing sentences annotated as *uncertain*, one interesting property is to look at sentence lengths (measured in tokens). One hypothesis is that uncertain sentences are in general longer. In Figure 3 we see that in general, for all three annotators, uncertain sentences are longer than certain sentences. This result is, of course, highly influenced by the skewness of the data (i.e. uncertain sentences are in minority), but it is clear that uncertain sentences, in general, are longer on average. It is interesting to note that the annotator SLS has, in most cases, annotated longer sentences as uncertain, compared to UCS and ULC. Moreover, *geriatrics*, with relatively high IAA but relatively low amounts of uncertain sentences, has well above average sentence lengths in the *uncertain* class.

4.2 Token level annotations

When it comes to the token level annotations, *speculative words* and *negations*, we observed very high IAA for *negations* (0.95 F_1-score (exact match) on average in the full corpus, the lowest for *neurology*, 0.94). These annotations were highly lexical (13 unique tokens) and unambiguous, and spread evenly across the two sentence level annotation classes (ranging between 1 and 3 percent of the total amount of tokens per class). Moreover, all negations were unigrams.

On the other hand, we observed large variations in IAA results for *speculative words*. In Figure 4, we see that there are considerable differences between exact and partial matches[4] between all annotator pairs, indicating individual differences in the interpretations of what constitutes a speculative word and how many tokens they cover, and the lexicality is not as evident as for negations. The highest level of agreement we find between UCS/ULC in *orthopaedics* (0.65 F_1-score, partial match) and *neurology* (0.64 F_1-score, partial match), and the lowest in *infection* (UCS/SLS, 0.31 F_1-score).

[4]Partial matches are measured on a character level.

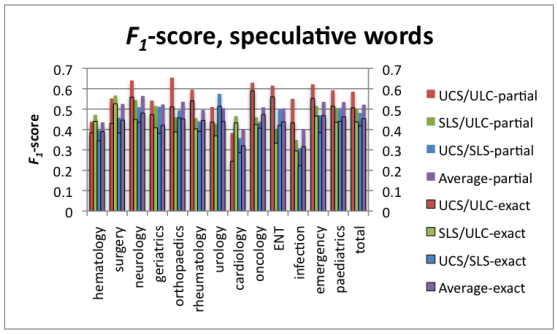

Figure 4: F_1-score, speculative words, exact and partial match.

4.2.1 Speculative words – most common

The low IAA results for *speculative words* invites a deeper analysis for this class. How is this interpreted by the individual annotators? First, we look at the most common tokens annotated as *speculative words*, shared by the three annotators: ”?”, ”sannolikt” (*likely*), ”ev” (*possibly*, abbreviated), ”om” (*if*). The most common speculative words are all unigrams, for all three annotators. These tokens are similar to the most common speculative words in the clinical BioScope subcorpus, where *if*, *may* and *likely* are among the top five most common. Those tokens that are most common per annotator and not shared by the other two (among the five most frequent) include ”bedöms” (*judged*), ”kan” (*could*), ”helt” (*completely*) and ”ställningstagande” (*standpoint*).

Looking at *neurology* and *urology*, with a higher overall average amount of uncertain sentences, we find that the most common words for *neurology* are similar to those most common in total, while for *urology* we find more n-grams. In Table 3, the five most common speculative words per annotator for neurology and urology are presented.

When it comes to the unigrams, many of these are also *not* annotated as speculative words. For instance, ”om” (*if*), is annotated as speculative in only 9 percent on average of its occurrence in the neurological data (the same distribution holds, on average, in the total set). In Morante and Daelemans (2009), *if* is also one of the words that are subject to the majority of false positives in their automatic classifier. On the other hand, ”sannolikt” (*likely*) is almost always annotated as a speculative word (over 90 percent of the time).

	UCS	ULC	SLS
neurology	? sannolikt (*likely*) kan (*could*) om (*if*) pröva (*try*) ter (*seem*)	? kan (*could*) sannolikt (*likely*) om (*if*) verkar (*seems*) ev (*possibly*, abbr)	? sannolikt (*likely*) ev (*possibly*, abbr) om (*if*) ställningstagande (*standpoint*) möjligen (*possibly*)
urology	kan vara (*could be*) tyder på(*indicates*) ev (*possibly*, abbr) misstänkt (*suspected*) kanske (*perhaps*) planeras tydligen (*apparently planned*)	mycket (*very*) inga tecken (*no signs*) kan vara (*could be*) kan (*could*) tyder (*indicates*) misstänkt (*suspected*)	tyder på(*indicates*) i första hand (*primarily*) misstänkt (*suspected*) kanske (*perhaps*) skall vi försöka (*should we try*) kan vara (*could be*)

Table 3: Most common speculative words per annotator for *neurology* and *urology*.

4.2.2 Speculative words – n-grams

Speculative words are, in Swedish clinical text, clearly not simple lexical unigrams. In Figure 5 we see that the average length of tokens annotated as *speculative words* is, on average, 1.34, with the longest in *orthopaedics* (1.49) and *urology* (1.46). We also see that SLS has, on average, annotated longer sequences of tokens as *speculative words* compared to UCS and ULC. The longest n-grams range between three and six tokens, e.g. "kan inte se några tydliga" (*can't see any clear*), "kan röra sig om" (*could be about*), "inte helt har kunnat uteslutas" (*has not been able to completely exclude*), "i första hand" (*primarily*). In many of these cases, the strongest indicator is actually a unigram ("kan" (*could*)), within a verb phrase. Moreover, negations inside a *speculative word* annotation, such as "inga tecken" (*no signs*) are annotated differently among the individual annotators.

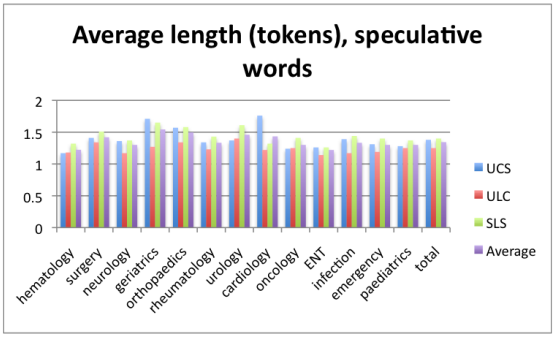

Figure 5: Average length, speculative words.

4.3 Examples

We have observed low average pairwise IAA for sentence level annotations in the *uncertain* class, with more or less large differences between the annotator pairs. Moreover, at the token level and for the class *speculative words*, we also see low average agreement, and indications that *speculative words* often are n-grams. We focus on the clinical practices *neurology*, because of its average large proportion of uncertain sentences, *geriatrics* for its high IAA results for UCS/ULC and low average proportion of uncertain sentences, and finally *surgery*, for its large discrepancy in proportions and low average IAA results.

In Example 1 we see a sentence where two annotators (ULC, SLS) have marked the sentence as *uncertain*, also marking a unigram ("ospecifik" (*unspecific*) as a *speculative word*. This example is interesting since the utterance is ambiguous, it can be judged as certain as in *the dizziness is confirmed to be of an unspecific type* or uncertain as in *the type of dizziness is unclear*, a type of utterance which should be clearly addressed in the guidelines.

<C> Yrsel av ospecifik typ. </C>

<U> Yrsel av <S> ospecifik </S> typ. </U>

<U> Yrsel av <S> ospecifik </S> typ. </U>

Dizziness of unspecific type

Example 1: Annotation example, *neurology*. Ambiguous sentence, *unspecific* as a possible speculation cue. C = Certain, U = Uncertain, S = Speculative words.

An example of different interpretations of the minimum span a *speculative word* covers is given in Example 2. Here, we see that "inga egentliga märkbara" (*no real apparent*) has been annotated in three different ways. It is also interesting to

note the role of the negation as part of amplifying speculation. Several such instances were marked by the annotators (for further examples, see Dalianis and Velupillai (2010)), which conforms well with the findings reported in Kilicoglu and Bergler (2008), where it is showed that explicit certainty markers together with negation are indicators of speculative language. In the BioScope corpus (Vincze et al., 2008), such instances are marked as speculation cues. This example, as well as Example 1, is also interesting as they both clearly are part of a longer passage of reasoning of a patient, with no particular diagnosis mentioned in the current sentence. Instead of randomly extracting sentences from the free text entry *Assessment*, one possibility would be to let the annotators judge all sentences in an entry (or a full EHR). Doing this, differences in where speculative language often occur in an EHR (entry) might become evident, as for scientific writings, where it has been showed that speculative sentences occur towards the end of abstracts (Light et al., 2004).

<U> <S><N> Inga </N> egentliga </S> <S> märkbara</S> minnessvårigheter under samtal. </U>.

<U> <N> Inga </N> <S> egentliga </S> märkbara minnessvårigheter under samtal. </U>.

<U> <S><N> Inga </N> egentliga märkbara </S> minnessvårigheter under samtal. </U>.

No real apparent memory difficulties during conversation

Example 2: Annotation example, *neurology*. Different annotation coverage over negation and speculation. C = Certain, U = Uncertain, S = Speculative words, N = Negation

In *geriatrics*, we have observed a lower than average amount of uncertain sentences, and high IAA between UCS and ULC. In Example 3 we see a sentence where UCS and ULC have matching annotations, whereas SLS has judged this sentence as certain. This example shows the difficulty of interpreting expressions indicating possible speculation – is ”ganska” (*relatively*) used here as a marker of certainty (as certain as one gets when diagnosing this type of illness)?

The word ”sannolikt” (*likely*) is one of the most common words annotated as a speculative word in the total corpus. In Example 4, we see a sentence where the annotators UCS and SLS have judged it to be *uncertain*, while UCS and ULC have marked the word ”sannolikt” (*likely*) as a *speculative word*. This is an interesting example, through informal discussions with clinicians we were informed that this word might as well be used as a marker of high certainty. Such instances show the need for using domain experts in future annotations of similar corpora.

<U> Både anamnestiskt och testmässigt <S> ganska </S> stabil vad det gäller Alzheimer sjukdom. </U>.

<U> Både anamnestiskt och testmässigt <S> ganska </S> stabil vad det gl̈er Alzheimer sjukdom. </U>.

<C> Både anamnestiskt och testmässigt ganska stabil vad det gäller Alzheimer sjukdom. </C>.

Both anamnesis and tests relatively stabile when it comes to Alzheimer's disease.

Example 3: Annotation example, *geriatrics*. Different judgements for the word ”ganska” (*relatively*). C = Certain, U = Uncertain, S = Speculative words.

<C>En 66-årig kvinna med <S>sannolikt</S> 2 synkrona tumörer vänster colon/sigmoideum och där till levermetastaser.</C>.

<U>En 66-årig kvinna med <S>sannolikt</S> 2 synkrona tumörer vänster colon/sigmoideum och där till levermetastaser.</U>.

<C>En 66-årig kvinna med sannolikt 2 synkrona tumörer vänster colon/sigmoideum och där till levermetastaser.</C>.

A 66 year old woman likely with 2 synchronous tumours left colon/sigmoideum in addition to liver metastasis.

Example 4: Annotation example, *surgery*. Different judgements for the word ”sannolikt” (*likely*). C = Certain, U = Uncertain, S = Speculative words.

5 Discussion

We have presented an analysis of an initial annotation trial for the identification of uncertain sentences as well as for token level cues (*speculative words*) across different clinical practices. Our main findings are that IAA results for both sentence level annotations of uncertainty and token level annotations for speculative words are, on av-

erage, fairly low, with higher average agreement in *geriatrics* and *rheumatology* (see Figures 1 and 2). Moreover, by analyzing the individual distributions for the classes *uncertain* and *speculative words*, we find that *neurology* has the highest average amount of uncertain sentences, and *cardiology* the lowest. On average, the amount of uncertain sentences ranges between 9 and 12 percent, which is in line with previous work on sentence level annotations of uncertainty (see Section 2).

We have also showed that the most common *speculative words* are unigrams, but that a substantial amount are n-grams. The n-grams are, however, often part of verb phrases, where the head is often the speculation cue. However, it is evident that speculative words are not always simple lexical units, i.e. syntactic information is potentially very useful. Question marks are the most common entities annotated as *speculative words*. Although these are not interesting indicators in themselves, it is interesting to note that they are very common in clinical documentation.

From the relatively low IAA results we draw the conclusion that this task is difficult and requires more clearly defined guidelines. Moreover, using *naive* coders on clinical documentation is possibly not very useful if the resulting annotations are to be used in, e.g. a Text Mining application for medical researchers. Clinical documentation is highly domain-specific and contains a large amount of internal jargon, which requires judgements from clinicians. However, we find it interesting to note that we have identified differences between different clinical practices. A consensus corpus has been created from the resulting annotations, which has been used in an experiment for automatic classification, see Dalianis and Skeppstedt (2010) for initial results and evaluation.

During discussions among the annotators, some specific problems were noted. For instance, the extracted sentences were not always about the patient or the current status or diagnosis, and in many cases an expression could describe (un)certainty of someone other than the author (e.g. another physician or a family member), introducing aspects of perspective. The sentences annotated as *certain*, are difficult to interpret, as they are simply *not uncertain*. We believe that it is important to introduce further dimensions, e.g. explicit certainty, and focus (*what* is (un)certain?), as well as time (e.g. *current* or *past*).

6 Conclusions

To our knowledge, there is no previous research on annotating Swedish clinical text for sentence and token level uncertainty together with an analysis of the differences between different clinical practices. Although the initial IAA results are in general relatively low for all clinical practice groups, we have identified indications that *neurology* is a practice which has an above average amount of uncertain elements, and that *geriatrics* has a below average amount, as well as higher IAA. Both these disciplines would be interesting to continue the work on identifying speculative language.

It is evident that clinical language contains a relatively high amount of uncertain elements, but it is also clear that naive coders are not optimal to use for interpreting the contents of EHRs. Moreover, more care needs to be taken in the extraction of sentences to be annotated, in order to ensure that the sentences actually describe reasoning about the patient status and diagnosis. For instance, instead of randomly extracting sentences from within a free text entry, it might be better to let the annotators judge all sentences within an entry. This would also enable an analysis of whether speculative language is more or less frequent in specific parts of EHRs.

From our findings, we plan to further develop the guidelines and particularly focus on specifying the minimal entities that should be annotated as *speculative words* (e.g. ”kan” (*could*)). We also plan to introduce further levels of dimensionality in the annotation task, e.g. cues that indicate a high level of certainty, and to use domain experts as annotators. Although there are problematic issues regarding the use of *naive* coders for this task, we believe that our analysis has revealed some properties of speculative language in clinical text which enables us to develop a useful resource for further research in the area of speculative language. Judging an instance as being certain or uncertain is, perhaps, a task which can never exclude subjective interpretations. One interesting way of exploiting this fact would be to exploit individual annotations similar to the work presented in Reidsma and op den Akker (2008). Once we have finalized the annotated set, and ensured that no identifiable information is included, we plan to make this resource available for further research.

References

Hercules Dalianis and Maria Skeppstedt. 2010. Creating and Evaluating a Consensus for Negated and Speculative Words in a Swedish Clinical Corpus. To be published in the proceedings of the Negation and Speculation in Natural Language Processing Workshop, July 10, Uppsala, Sweden.

Hercules Dalianis and Sumithra Velupillai. 2010. How Certain are Clinical Assessments? Annotating Swedish Clinical Text for (Un)certainties, Speculations and Negations. In *Proceedings of the of the Seventh International Conference on Language Resources and Evaluation, LREC 2010*, Valletta, Malta, May 19-21.

J. L. Hobby, B. D. M. Tom, C. Todd, P. W. P. Bearcroft, and A. K. Dixon. 2000. Communication of doubt and certainty in radiological reports. *The British Journal of Radiology*, 73:999–1001, September.

R. Khorasani, D. W. Bates, S. Teeger, J. M. Rotschild, D. F. Adams, and S. E. Seltzer. 2003. Is terminology used effectively to convey diagnostic certainty in radiology reports? *Academic Radiology*, 10:685–688.

Halil Kilicoglu and Sabine Bergler. 2008. Recognizing speculative language in biomedical research articles: a linguistically motivated perspective. *BMC Bioinformatics*, 9(S-11).

Marc Light, Xin Ying Qiu, and Padmini Srinivasan. 2004. The language of bioscience: Facts, speculations, and statements in between. In Lynette Hirschman and James Pustejovsky, editors, *HLT-NAACL 2004 Workshop: BioLINK 2004, Linking Biological Literature, Ontologies and Databases*, pages 17–24, Boston, Massachusetts, USA, May 6. Association for Computational Linguistics.

Roser Morante and Walter Daelemans. 2009. Learning the scope of hedge cues in biomedical texts. In *BioNLP '09: Proceedings of the Workshop on BioNLP*, pages 28–36, Morristown, NJ, USA. Association for Computational Linguistics.

Philip V. Ogren. 2006. Knowtator: a protégé plug-in for annotated corpus construction. In *Proceedings of the 2006 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, pages 273–275, Morristown, NJ, USA. Association for Computational Linguistics.

Arzucan Özgür and Dragomir R. Radev. 2009. Detecting speculations and their scopes in scientific text. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 1398–1407, Singapore, August. Association for Computational Linguistics.

Dennis Reidsma and Rieks op den Akker. 2008. Exploiting 'subjective' annotations. In *HumanJudge '08: Proceedings of the Workshop on Human Judgements in Computational Linguistics*, pages 8–16, Morristown, NJ, USA. Association for Computational Linguistics.

Victoria L. Rubin, Elizabeth D. Liddy, and Noriko Kando. 2006. Certainty identification in texts: Categorization model and manual tagging results. In *Computing Affect and Attitutde in Text: Theory and Applications*. Springer.

Veronika Vincze, György Szarvas, Richárd Farkas, György Móra, and János Csirik. 2008. The bioscope corpus: biomedical texts annotated for uncertainty, negation and their scopes. *BMC Bioinformatics*, 9(S-11).

J. W. Wilbur, A. Rzhetsky, and H. Shatkay. 2006. New directions in biomedical text annotation: definitions, guidelines and corpus construction. *BMC Bioinformatics*, 7:356+, July.

Does Negation Really Matter?

Ira Goldstein
University at Albany, SUNY
Albany, NY USA
ig4895@albany.edu

Özlem Uzuner
University at Albany, SUNY
Albany, NY USA
ouzuner@albany.edu

Abstract

We explore the role negation and speculation identification plays in the multi-label document-level classification of medical reports for diseases. We identify the polarity of assertions made on noun phrases which reference diseases in the medical reports. We experiment with two machine learning classifiers: one based upon Lucene and the other based upon BoosTexter. We find the performance of these systems on document-level classification of medical reports for diseases fails to show improvement when their input is enhanced by the polarity of assertions made on noun phrases. We conclude that due to the nature of our machine learning classifiers, information on the polarity of phrase-level assertions does not improve performance on our data in a multi-label document-level classification task.

1 Introduction

In the medical domain, a substantial amount of patient data is stored as free text in patient medical report narratives (Spat et al. 2008) and needs to be processed in order to be converted to more widely-useful structured information. These narratives contain a variety of useful information that can support syndromic surveillance (Shapiro 2004), decision support (Fiszman et al. 2000), and problem list generation (Sibanda et al. 2006).

Physicians often assert negative or speculative diagnoses in medical reports (Rao et al. 2003) to keep track of all potential diagnoses that have been considered and to provide information that contrasts with the positive diagnoses (Kim and Park 2006). The noun phrases (NP) associated with negative and speculative assertions in medical reports may be confused with positively asserted NPs, thereby adversely affecting automated classification system performance. In the medical domain, verbs often play a reduced role or are implied in assertions. We therefore focus our investigation of assertions on NPs.

In this paper, we describe the polarity of an assertion as being positive, speculative, or negative. Assertion classification is a generally accepted means for resolving problems caused by negation and speculation. Averbuch et al. (2004) use context to identify negative/positive instances of various symptoms. Mutalik et al. (2001) show that the Unified Medical Language System (UMLS) Metathesaurus can be used to reliably detect negated concepts in medical narratives. Harkema et al. (2009) develop ConText to determine not only positive and negative assertions, but also assertions referencing someone other than the patient.

The literature is filled with reports of systems which employ assertion classification (e.g., Google Scholar lists 134 documents citing Chapman et al.'s (2001) NegEx). However, few reports describe how much assertion classification contributes to the final system performance. Two exceptions are Goldstein et al. (2007) and Ambert and Cohen (2009).

Goldstein et al. develop a hand-crafted rule based system to classify radiological reports from the 2007 Computational Medicine Center (CMC) Challenge (Pestian et al. 2007). They show that negation and speculation play key roles in classifying their reports. Ambert and Cohen apply a machine learning (ML) approach to classifying discharge summaries from the 2008 i2b2 Obesity Challenge (Uzuner 2008). They report that due to "false negations," simply adding negation detection to their base system does not consistently improve performance. Prompted by these contradicting results in the literature, we explore the role assertion classification plays in the multi-label classification of medical reports from both the CMC and i2b2 challenges.

We attempt to improve document-level classification performance of two multi-label ML classifiers by identifying the polarity of assertions on NPs. We experiment with medical reports from two different corpora. We detect NPs which reference diseases. We then identify the polarity of the assertion made for each NP. We show that enriching reports with the polarity of the assertions does not improve performance for multi-label document-level classification of medical

Proceedings of the Workshop on Negation and Speculation in Natural Language Processing, pages 23–27,
Uppsala, July 2010.

reports into diseases in our corpora. Our findings imply that, despite common practice, the contribution of assertion classification may be limited when employing ML approaches to predicting document-level labels of medical reports.

2 Data

The data were provided by the CMC challenge (Pestian et al. 2007) and the i2b2 Obesity Challenge (Uzuner 2008). Both data sets had been de-identified (anonymized) and, where appropriate, re-identified with surrogates. Our task is to determine the presence of diseases in the patient based upon medical report narratives. The institutional review boards of the SUNY Albany and Partners HealthCare approved this study.

2.1 CMC Data Set

The CMC data set consists of a training set of 978 radiology reports and a test set of 976 radiology reports. Each report is labeled with ICD-9-CM (National Center for Health Statistics 2010) standard diagnostic classification codes.

The reports have been hand labeled with 45 ICD-9-CM. Each code represents a distinct disease present in the patient. The codes reflect only the definite diagnoses mentioned in that report. At least one code is assigned to each report. Multiple codes per report are allowed. For each report in the test set, we predict which diseases are present in the patient and label the report with the ICD-9-CM code for that disease. Any code not assigned to a report implies that the corresponding disease is not present in the patient.

2.2 i2b2 Data Set

The i2b2 data set consists of a training set of 720 discharge summaries and a test set of 501 discharge summaries. These medical reports range in size from 133 words to more than 3000 words. The reports have been labeled for information on obesity and 15 of its most frequent co-morbidities. For each report, each disease is labeled as being present, absent, or questionable in the patient, or unmentioned in the narrative. Multiple codes per report are allowed.

Since we are interested in those diseases present in the patient, we retain the present class and collapse the absent, questionable, and unmentioned categories into a not present class. For each report in the test set we predict whether each of the 16 diseases is present or not present in the patient. We label each report with our prediction for each of the 16 diseases.

3 Methods

We preprocess the medical report narratives with a Noun Phrase Detection Pre-processor (NPDP) to detect noun phrases referencing diseases. We implement our own version of ConText (Harkema et al. 2009), enhance it to also detect speculation, and employ it to identify the polarity of assertions made on the detected NPs. We expand the text of the medical reports with asserted NPs. We conflate lexical variations of words. We train two different types of classifiers on each of the training sets. We apply labels to both the expanded and non-expanded reports using two ML classifiers. We evaluate and report results only on the test sets.

3.1 Noun Phrase and Assertion Detection

We detect noun phrases via an NPDP. We build our NPDP based on MetaMap (Aronson 2001). The NPDP identifies NPs which reference diseases in medical reports. We select 17 UMLS semantic types whose concepts can assist in the classification of diseases. First, NPDP maps NPs in the text to UMLS semantic types. If the mapped semantic type is one of the target semantic types, NPDP then tags the NP.

NPDP uses the pre-UMLS negation phrases of Extended NegEx (Sibanda et al. 2006) to identify adjectives indicating the absence or uncertainty of each tagged NPs. It differentiates these adjectives from all other adjectives modifying tagged NPs. For example, *possible* in *possible reflux* is excluded from the tagged NP, whereas *severe* in *severe reflux* is retained. We then identify the polarity of the assertion made on each NP. In order to distinguish the polarity of the assertions from one another, we do not modify the positive assertions, but transform the negative and speculative assertions in the following manner: Sentences containing negative assertions are repeated and modified with the NP pre-pended with “abs” (e.g., “Patient denies fever.” is repeated as “Patient denies absfever.”). Similarly, sentences containing speculative assertions are repeated and modified with the NP pre-pended with “poss”. We refer to these transformed terms as *asserted noun phrases*. We assert NPs for the unmodified text of both the data sets. Table 1 provides a breakdown of the assertions for each of the detected NPs for each of the data sets.

We examine the performance of our enhanced implementation of ConText by comparing its results against CMC test set NPs manually annotated by a nurse librarian and author IG. Table 2

shows the performance for each of the three polarities. We find these results to be comparable to those reported in the literature: Mutalik et al.'s (2001) NegFinder finds negated concepts with a recall of .957; Chapman et al.'s (2001) NegEx report a precision of .8449 and a recall of .8241.

Assertion	CMC		i2b2	
	Training	Test	Training	Test
Positive	2,168	2,117	47,860	34,112
Speculative	312	235	3,264	2,166
Negative	351	353	8,202	5,654

Table 1 - Distribution of Asserted Noun Phrases for both the CMC and i2b2 data sets.

Assertion	Precision	Recall	F1-Measure
Positive	0.991	0.967	0.979
Speculative	0.982	0.946	0.964
Negative	0.770	0.983	0.864

Table 2 - Assertion Performance on the CMC test set.

3.2 Lucene Classifier

We follow the k-Nearest Neighbor (Cover and Hart 1967) process previously described in Goldstein et al. (2007) to build our Lucene-based classifier. Classification is based on the nearest training samples, as determined by the feature vectors. This approach assumes that similar training samples will cluster together in the feature vector space. The nearest training samples are considered to be those that are most similar to the data sample.

We build our Lucene-based classifier using Apache Lucene (Gospodnetić and Hatcher 2005). We use the Lucene library to determine the similarity of medical report narratives. We determine which training reports are similar to the target report based upon their text. For each target report we retrieve the three most similar training reports and assign to the target report any codes that are used by the majority of these reports. In cases where the retrieved reports do not provide a majority code, the fourth nearest training report is used. If a majority code is still not found, a NULL code is assigned to the target report.

We first run the Lucene Classifier on lower case, stemmed text of the medical reports. We refer to this as the *Base Lucene Classifier* run. We next run the Lucene Classifier on the text expanded with asserted noun phrases. We refer to this as the *Asserted Lucene Classifier* run.

3.3 BoosTexter Classifier

BoosTexter (Schapire and Singer 2000) builds classifiers from textual data by performing multiple iterations of dividing the text into subsamples upon which weak decision-stub learners are trained. Among these weak learners, BoosTexter retains those that perform even marginally better than chance. After a set number of iterations, the retained weak learners are combined into the final classifier. BoosTexter classifies text using individual words (unigrams), strings of consecutive words (n-grams), or strings of non-consecutive words, without considering semantics.

We cross-validate BoosTexter (tenfold) on the CMC training set. We establish the optimal parameters on the CMC training set to be 1100 iterations, with n-grams of up to four words. We find the optimal parameters of the i2b2 training set to be similar to those of the CMC training set. For consistency, we apply the parameters of 1100 iterations and n-grams of up to four words to both data sets. In addition, we apply unigrams to BoosTexter in order to provide BoosTexter classifier results that are comparable to those of the Lucene classifiers.

We create two classifiers with BoosTexter using the lower case, stemmed text of the medical reports: one with unigrams and one with n-grams. We refer to these as *Base BoosTexter Classifier* runs. For each of unigrams and n-grams, we create runs on the text expanded with the asserted noun phrases. We refer to these as *Asserted BoosTexter Classifier* runs.

4 Evaluation

We evaluate our classifiers on both the plain text of the reports and on text expanded with asserted NPs. We present results in terms of micro-averaged precision, recall, and F1-measure (Özgür et al. 2005). We check the significance of classifier performance differences at α=0.10. We apply a two-tailed Z test, with $Z = \pm 1.645$.

5 Results and Discussion

Table 3 and Table 4 show our systems' performances. We predict ICD-9-CM codes for each of the 976 CMC test reports. We predict whether or not each of 16 diseases is present in the patient for each of the 501 i2b2 test set reports.

	Negative Reports			Positive Reports		
Run	Precision	Recall	F1-Measure	Precision	Recall	F1-Measure
CMC Base	0.991	0.993	0.992	0.717	0.664	0.690
CMC Asserted	0.991	0.992	0.992	0.712	0.668	0.690
i2b2 Base	0.905	0.886	0.896	0.612	0.660	0.635
i2b2 Asserted	0.904	0.890	0.897	0.618	0.651	0.634

Table 3 - Lucene Classifier's Performance.

The Asserted Lucene and BoosTexter Classifier runs show no significant difference in performance from their Base runs on either corpus. These results indicate that asserted noun phrases do not contribute to the document-level classification of our medical reports

5.1 Contribution of Asserted Noun Phrases

Through analysis of the Base and Asserted runs, we find enough similarities in the text of the training and test reports for a given class to allow our ML classifiers to correctly predict the labels without needing to identify the polarity of the assertions made on individual NPs. For example, for the CMC target report 97729923:

```
5-year-9-month - old female
with two month history of
cough. Evaluate for pneumonia.
No pneumonia.
```

the Base Lucene Classifier retrieves report 97653364:

```
Two - year-old female with
cough off and on for a month
(report states RSV nasal
wash).
No radiographic features of
pneumonia.
```

which allows the system to classify the target report with the ICD-9-CM code for cough. While identifying the polarity of the assertions for pneumonia strengthens the evidence for cough and not pneumonia, it cannot further improve the already correct document-level classification. These unenhanced assertions do not stand in the way of correct classification by our systems.

5.2 Approach, Data, and Task

Hand-crafted rule-based approaches usually encode the most salient information that the experts would find useful in classification and would therefore benefit from explicit assertion classification subsystems, e.g., Goldstein et al., (2007). On the other hand, ML approaches have the ability to identify previously undetected patterns in data (Mitchell et al. 1990). This enables ML approaches to find patterns that may not be obvious to experts, while still performing correct classification. Therefore, the contribution of asserted NPs appears to be limited when applied to ML approaches to document-level classification of medical reports. This is not to say that an ML approach to document-level classification will never benefit from identifying the polarity of NPs; only that on our data we find no improvement.

		Negative Reports			Positive Reports		
Run		Precision	Recall	F1-Measure	Precision	Recall	F1-Measure
CMC unigram	Base	0.993	0.995	0.994	0.812	0.747	0.778
	Asserted	0.993	0.996	0.995	0.837	0.767	0.800
CMC ngram	Base	0.995	0.996	0.996	0.865	0.812	0.838
	Asserted	0.995	0.996	0.996	0.866	0.812	0.839
i2b2 unigram	Base	0.970	0.973	0.917	0.902	0.889	0.895
	Asserted	0.970	0.975	0.973	0.908	0.891	0.899
i2b2 ngram	Base	0.971	0.976	0.974	0.911	0.895	0.903
	Asserted	0.974	0.977	0.975	0.914	0.903	0.908

Table 4 - BoosTexter Classifier's Performance.

The CMC and i2b2 data sets can each be described as being homogenous; they come from a relatively small communities and limited geographic areas. In these data, variation in vocabulary that might arise from the use of regional expressions would be limited. This would be especially true for the CMC data since it comes from a single medical department at a single hospital. It would not be surprising for colleagues in a given department who work together for a period of time to adopt similar writing styles and to employ consistent terminologies (Suchan 1995).

Our task is one of multi-label document-level classification. Working at the document level, each negative and speculative assertion would play only a small role in predicting class labels.

The homogeneity of the text in our data sets, and the task of document-level classification may have been factors in our results. Future research should examine how the characteristics of the data and the nature of the task affect the role of assertion classification.

6 Conclusion

Identifying the polarity of phrase-level assertions in document-level classification of medical reports may not always be necessary. The specific task and approach applied, along with the characteristics of the corpus under study, should be considered when deciding the appropriateness of assertion classification. The results of this study show that on our data and task, identifying the polarity of the assertions made on noun phrases does not improve machine learning approaches to multi-label document-level classification of medical reports.

References

Kyle H. Ambert and Aaron M. Cohen. 2009. A System for Classifying Disease Comorbidity Status from Medical Discharge Summaries Using Automated Hotspot and Negated Concept Detection. *Journal of the American Medical Informatics Association* 16(4):590-95.

Alan R. Aronson. 2001. Effective Mapping of Biomedical Text to the UMLS Metathesaurus: The Metamap Program. *Proceedings of the AMIA symposium*. 17-21.

Mordechai Averbuch, Tom H. Karson, Benjamin Ben-Ami, Oded Maimon, and Lior Rokach. 2004. Context-Sensitive Medical Information Retrieval. *Medinfo. MEDINFO* 11(Pt 1):282-86.

Wendy W. Chapman, Will Bridewell, Paul Hanbury, Gregory F. Cooper, and Bruce G. Buchanan. 2001. A Simple Algorithm for Identifying Negated Findings and Diseases in Discharge Summaries. *Journal of Biomedical Informatics* 34(5):301-10.

Thomas M. Cover and Peter E. Hart. 1967. Nearest Neighbor Pattern Classification. *IEEE Transactions on Information Theory* 13(1):21-27.

Marcelo Fiszman, Wendy W. Chapman, Dominik Aronsky, and R. Scott Evans. 2000. Automatic Detection of Acute Bacterial Pneumonia from Chest X-Ray Reports. *Journal of the American Medical Informatics Association* 7:593-604.

Ira Goldstein, Anna Arzumtsyan, and Özlem Uzuner. 2007. Three Approaches to Automatic Assignment of ICD-9-CM Codes to Radiology Reports. *Proceedings of the AMIA symposium*. 279-83.

Otis Gospodnetić and Erik Hatcher. 2005. *Lucene in Action*. Greenwich, CT: Manning.

Henk Harkema, John N. Dowling, Tyler Thornblade, and Wendy W. Chapman. 2009. Context: An Algorithm for Determining Negation, Experiencer, and Temporal Status from Clinical Reports. *Journal Of Biomedical Informatics* 42(5):839-51.

Jung-Jae Kim and Jong C. Park. 2006. Extracting Contrastive Information from Negation Patterns in Biomedical Literature. *ACM Transactions on Asian Language Information Processing (TALIP)* 5(1):44-60.

Tom Mitchell, Bruce Buchanan, Gerald DeJong, Thomas Dietterich, Paul Rosenbloom, and Alex Waibel. 1990. Machine Learning. *Annual Review of Computer Science. Vol.4*. Eds. Joseph F. Traub, Barbara J. Grosz, Butler W. Lampson and Nils J. Nilsson. Palo Alto, CA: Annual Reviews.

Pradeep G. Mutalik, Aniruddha Deshpande, and Prakash M. Nadkarni. 2001. Use of General-Purpose Negation Detection to Augment Concept Indexing of Medical Documents: A Quantitative Study Using the UMLS. *Journal of the American Medical Informatics Association* 8(6):598-609.

National Center for Health Statistics. 2010. *ICD - ICD-9-CM - International Classification of Diseases, Ninth Revision, Clinical Modification*. Accessed: May 1, 2010. <www.cdc.gov/nchs/icd/icd9cm.htm>.

Arzucan Özgür, Levent Özgür, and Tunga Güngör. 2005. Text Categorization with Class-Based and Corpus-Based Keyword Selection. *ISCIS 2005*. Eds. Pınar Yolum, Tunga Güngör, Fikret Gürgen and Can Özturan. Istanbul, Turkey: Springer. 606-15 of *Lecture Notes in Computer Science*.

John P. Pestian, Christopher Brew, Pawel Matykiewicz, D. J. Hovermale, Neil Johnson, K. Bretonnel Cohen, and Włodzisław Duch. 2007. A Shared Task Involving Multi-Label Classification of Clinical Free Text. *ACL:BioNLP*. Prague: Association for Computational Linguistics. 97-104.

R. Bharat Rao, Sathyakama Sandilya, Radu Stefan Niculescu, Colin Germond, and Harsha Rao. 2003. Clinical and Financial Outcomes Analysis with Existing Hospital Patient Records. *Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*: ACM Press New York, NY, USA. 416-25.

Robert E. Schapire and Yoram Singer. 2000. Boostexter: A Boosting-Based System for Text Categorization. *Machine Learning* 39(2):135-68.

Alan R. Shapiro. 2004. Taming Variability in Free Text: Application to Health Surveillance. *MMWR. Morbidity And Mortality Weekly Report* 53 Suppl:95-100.

Tawanda Carleton Sibanda, T. He, Peter Szolovits, and Özlem Uzuner. 2006. Syntactically-Informed Semantic Category Recognition in Discharge Summaries. *Proceedings of the AMIA symposium*. 714-8.

Stephan Spat, Bruno Cadonna, Ivo Rakovac, Christian Gütl, Hubert Leitner, Günther Stark, and Peter Beck. 2008. Enhanced Information Retrieval from Narrative German-Language Clinical Text Documents Using Automated Document Classification. *Studies In Health Technology And Informatics* 136:473-78.

Jim Suchan. 1995. The Influence of Organizational Metaphors on Writers' Communication Roles and Stylistic Choices. *Journal of Business Communication* 32(1):7-29.

Özlem Uzuner. 2008. Second I2b2 Workshop on Natural Language Processing Challenges for Clinical Records. *Proceedings of the AMIA symposium*:1252-53.

Speculation and negation annotation in natural language texts: what the case of BioScope might (not) reveal

Veronika Vincze
University of Szeged
Szeged, Hungary
vinczev@inf.u-szeged.hu

1 Introduction

In information extraction, it is of key importance to distinguish between facts and uncertain or negated information. In other words, IE applications have to treat sentences / clauses containing uncertain or negated information differently from factual information that is why the development of hedge and negation detection systems has received much interest – e.g. the objective of the CoNLL-2010 Shared Task was also to develop hedge detection systems (Farkas et al., 2010). For the training and evaluation of such systems, corpora annotated for negation and speculation are necessary.

There are several linguistic phenomena that can be grouped under the term uncertainty. Besides hedge and speculation, doubtful events are also considered as a subtype of uncertainty (Kim et al., 2008) and Ganter and Strube (2009) argue that the notion of *weasel words* are similar to hedges. A word is considered to be a weasel word if it creates an impression that something important has been said, but what is really communicated is vague, misleading, evasive or ambiguous, thus, it is also related to uncertainty. All these phenomena might be of interest for IE applications, which yields that the creation of corpora with uncertainty annotation is indispensable.

2 Related work

There exist some corpora that contain annotation for speculation and/or negation. The GENIA Event corpus (Kim et al., 2008) annotates biological events with negation and two types of uncertainty. In the BioInfer corpus (Pyysalo et al., 2007) biological relations are annotated for negation. The system developed by Medlock and Briscoe (2007) made use of a corpus consisting of six papers from genomics literature in which sentences were annotated for speculation. Settles et al. (2008) constructed a corpus where sentences are classified as either speculative or definite, however, no keywords are marked in the corpus and Shatkay et al. (2008) describe a database where sentences are annotated for certainty among other features. As a corpus specifically annotated for weasel words, WikiWeasel should be mentioned, which was constructed for the CoNLL-2010 Shared Task (Farkas et al., 2010) and contains Wikipedia paragraphs annotated for weasel words.

3 The BioScope corpus

The BioScope corpus (Vincze et al., 2008) is – to our best knowledge – the largest corpus available that is annotated for both negation and hedge keywords and the only one that contains annotation for linguistic scopes. It includes three types of texts from the biomedical domain – namely, radiological reports, biological full papers and abstracts from the GENIA corpus. (15 new full biomedical papers were annotated for hedge cues and their scopes, which served as the evaluation database of the CoNLL-2010 Shared Task (Farkas et al., 2010), and this dataset will be added to BioScope in the near future.) The annotation was carried out by two students of linguistics supervised by a linguist. Problematic cases were continuously discussed among the annotators and dissimilar annotations were later resolved by the linguist.

3.1 Annotation principles

In BioScope, speculation is understood as the possible existence of a thing is claimed – neither its existence nor its non-existence is known for sure. Only one level of uncertainty is marked (as opposed to the GENIA corpus (Kim et al., 2008) or Shatkay et al. (2008)) and no weasels are annotated. Negation is seen as the implication of non-existence of something.

The annotation was based on four basic principles:

Proceedings of the Workshop on Negation and Speculation in Natural Language Processing, pages 28–31,
Uppsala, July 2010.

- Each keyword has a scope.
- The scope must include its keyword.
- Min-max strategy:
  - The minimal unit expressing hedge/negation is marked as keyword.
  - The scope is extended to the maximal syntactic unit.
- No intersecting scopes are allowed.

These principles were determined at the very beginning of the annotation process and they were strictly followed throughout the corpus building.

3.2 Problematic cases

However, in some cases, some language phenomena seemed to contradict the above principles. These issues required a thorough consideration of the possible solutions in accordance with the basic principles in order to keep the annotation of the corpus as consistent as possible. The most notable examples include the following:

- Negative keywords without scope:

  > [Negative] chest radiograph.

  In this case, the scope contains only the keyword.

- Elliptic sentences

  > Moreover, ANG II stimulated NF-kappaB activation in human monocytes, but [not] in lymphocytes from the same preparation.

  With the present encoding scheme of scopes, there is no way to signal that the negation should be extended to the verb and the object as well.

- Nested scopes

  One scope includes another one:

  > These observations (suggest that TNF and PMA do (not lead to NF-kappa B activation through induction of changes in the cell redox status)).

  The semantic interpretation of such nested scopes should be understood as ”it is possible that there is no such an event that. . . ”.

- Elements in between keyword and target word

  Although *however* is not affected by the hedge cue in the following example, it is included in the scope since consecutive text spans are annotated as scopes:

  > (Atelectasis in the right mid zone is, however, <possible>).

- Complex keywords

  Sometimes a hedge / negation is expressed via a phrase rather than a single word: these are marked as complex keywords.

- Inclusion of modifiers and adjuncts

  It is often hard to decide whether a modifier or adjunct belongs to the scope or not. In order not to lose potentially important information, the widest scope possible is marked in each case.

- Intersecting scopes

  When two keywords occur within one sentence, their scopes might intersect, yielding one apparently empty scope (i.e. scope without keyword) and a scope with two keywords:

  > (Repression did ([not] <seem> to involve another factor whose activity is affected by the NSAIDs)).

  In such cases, one of the scopes (usually the negative one) was extended:

  > ((Repression did [not] <seem> to involve another factor whose activity is affected by the NSAIDs)).

On the other hand, there were some cases where the difficulty of annotation could be traced back to lexical issues. Some of the keyword candidates have several senses (e.g. *if*) or can be used in different grammatical structures (e.g. *indicate* vs. *indicate that*) and not all of them are to be marked as a keyword in the corpus. Thus, senses / usages to be annotated and those not to be annotated had to be determined precisely.

Finally, sometimes an apparently negative keyword formed part of a complex hedge keyword (e.g. *cannot be excluded*), which refers to the fact that speculation can be expressed also by a negated word, thus, the presence of a negative word does not automatically entail that the sentence is negated.

4 Outlook: Comparison with other corpora

Besides BioScope, the GENIA Event corpus (Kim et al., 2008) also contains annotation for negation and speculation. In order to see what the main differences are between the corpora, the annotation principles were contrasted:

- in GENIA Event, no modifier keywords are marked, however, in BioScope, they are;
- the scope of speculation and negation is explicitly marked in BioScope and it can be extended to various constituents within the clause / sentence though in GENIA Event, it is the event itself that is within the scope;
- two subtypes of uncertainty are distinguished in GENIA Event: *doubtful* and *probable*, however, in BioScope there is one umbrella term for them (*speculation*).

An essential difference in annotation principles between the two corpora is that GENIA Event follows the principles of event-centered annotation while BioScope annotation does not put special emphasis on events. Event-centered annotation means that annotators are required to identify as many events as possible within the sentence then label each separately for negation / speculation.

The multiplicity of events in GENIA and the maximum scope principle exploited in BioScope (see 3.1) taken together often yields that a GENIA event falls within the scope of a BioScope keyword, however, it should not be seen as a speculated or negated event on its own. Here we provide an illustrative example:

> In summary, our data suggest that changes in the composition of transcription factor AP-1 is a key molecular mechanism for increasing IL-2 transcription and may underlie the phenomenon of costimulation by EC.

According to the BioScope analysis of the sentence, the scope of *suggest* extends to the end of the sentence. It entails that in GENIA it is only the events *is a key molecular mechanism* and *underlie the phenomenon* that are marked as probable, nevertheless, the events *changes*, *increasing*, *transcription* and *costimulation* are also included in the BioScope speculative scope. Thus, within this sentence, there are six GENIA events out of which two are labeled as probable, however, in BioScope, all six are within a speculative scope.

In some cases, there is a difference in between what is seen as speculative / negated in the corpora. For instance, negated "investigation" verbs in Present Perfect are seen as doubtful events in GENIA and as negative events in BioScope:

> However, a role for NF-kappaB in human CD34(+) bone marrow cells has not been described.

According to GENIA annotation principles, the role has not been described, therefore it is doubtful what the role exactly is. However, in BioScope, the interpretation of the sentence is that there has not been such an event that the role for NF-kappaB in human CD34(+) bone marrow cells has been described. Thus, it is marked as negative.

Another difference between the annotation schemes of BioScope and GENIA is that instances of weaseling are annotated as probable events in GENIA, however, in BioScope they are not. An example for a weasel sentence is shown below:

> Receptors for leukocyte chemoattractants, including chemokines, are traditionally considered to be responsible for the activation of special leukocyte functions such as chemotaxis, degranulation, and the release of superoxide anions.

5 Conclusions

Some interesting conclusions can be drawn from the difficulties encountered during annotation process of the BioScope corpus. As for methodology, it is unquestionable that precisely defined rules (on scope marking, keyword marking and on the interpretation of speculation / negation) are essential for consistent annotation, thus, pre-defined guidelines can help annotation work a lot. However, difficulties or ambiguities not seen previously may emerge (and they really do) only during the process of annotation. In this way, a continuous reformulation and extension of annotation rules is required based on the corpus data. On the other hand, problematic issues sometimes might be solved in several different ways. When deciding on their final treatment, an ideal balance between gain and loss should be reached, in other words, the min-max strategy as a basic annotation

principle can also be applied here (minimize the loss and maximize the gain that the solution can provide).

Acknowledgments

This work was supported in part by the National Office for Research and Technology (NKTH, `http://www.nkth.gov.hu/`) of the Hungarian government within the framework of the project MASZEKER.

References

Richárd Farkas, Veronika Vincze, György Móra, János Csirik, and György Szarvas. 2010. The CoNLL-2010 Shared Task: Learning to Detect Hedges and their Scope in Natural Language Text. In *Proceedings of the Fourteenth Conference on Computational Natural Language Learning (CoNLL-2010): Shared Task*, pages 1–12, Uppsala, Sweden, July. Association for Computational Linguistics.

Viola Ganter and Michael Strube. 2009. Finding Hedges by Chasing Weasels: Hedge Detection Using Wikipedia Tags and Shallow Linguistic Features. In *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*, pages 173–176, Suntec, Singapore, August. Association for Computational Linguistics.

Jin-Dong Kim, Tomoko Ohta, and Jun'ichi Tsujii. 2008. Corpus annotation for mining biomedical events from literature. *BMC Bioinformatics*, 9(Suppl 10).

Ben Medlock and Ted Briscoe. 2007. Weakly Supervised Learning for Hedge Classification in Scientific Literature. In *Proceedings of the ACL*, pages 992–999, Prague, Czech Republic, June.

Sampo Pyysalo, Filip Ginter, Juho Heimonen, Jari Björne, Jorma Boberg, Jouni Järvinen, and Tapio Salakoski. 2007. BioInfer: a corpus for information extraction in the biomedical domain. *BMC Bioinformatics*, 8(50).

Burr Settles, Mark Craven, and Lewis Friedland. 2008. Active learning with real annotation costs. In *Proceedings of the NIPS Workshop on Cost-Sensitive Learning*, pages 1–10.

Hagit Shatkay, Fengxia Pan, Andrey Rzhetsky, and W. John Wilbur. 2008. Multi-dimensional classification of biomedical text: Toward automated, practical provision of high-utility text to diverse users. *Bioinformatics*, 24(18):2086–2093.

Veronika Vincze, György Szarvas, Richárd Farkas, György Móra, and János Csirik. 2008. The BioScope Corpus: Biomedical Texts Annotated for Uncertainty, Negation and their Scopes. *BMC Bioinformatics*, 9(Suppl 11):S9.

Automatic annotation of speculation in biomedical texts: new perspectives and large-scale evaluation

Julien Desclés **Olfa Makkaoui** **Taouise Hacène**
julien.descles@gmail.com olfa_makkaoui@yahoo.fr taouise@gmail.com
LaLIC Université Paris-Sorbonne Maison de la Recherche 28 rue Serpente, 75006 Paris

Abstract

One emergent field in text mining tools applied to biological texts is the automatic detection of speculative sentences. In this paper, we test on a large scale BioExcom, a rule-based system which annotates and categorizes automatically speculative sentences ("prior" and "new"). This work enables us to highlight a more restrictive way to consider speculations, viewed as a source of knowledge, and to discuss the criteria used to determine if a sentence is speculative or not. By doing so, we demonstrate the efficiency of BioExcom to extract these types of speculations and we argue the importance of this tool for biologists, who are also interested in finding hypotheses.

1 Introduction

In one of the first studies about biological speculations in Natural Language Processing, Light et al. (2004) have reported that biologists can have a dual posture concerning this subject:

> "*In the context of information retrieval, an example information need might be "I am looking for speculations about the X gene in liver tissue." One of the authors spoke at a research department of a drug company and the biologists present expressed this sort of information need. On the other hand, one of the authors has also encountered the opposite need: "I am looking for definite statements about transcription factors that interact with NF Kappa B." Both these information needs would be easier to fulfill if automated annotation of speculative passages was possible*." (Light et al., 2004)

In this quotation, the term "*speculation*" has not exactly the same meaning in these two statements (depending if it comes from computational linguists or from biologists). Indeed, because biologists are almost interested in knowing only factual statements, the information extraction tools have to remove all uncertain statements, identified as hedging or speculation, or at least to present them separately from definite results. Consequently, the vast majority of Natural Language Processing tools dealing with the identification of speculation have followed this very large meaning of the word "*speculation*", in order to avoid extracting uncertain information as factual information (Kilicoglu and Bergler, 2008; Medlock, 2008; Szarvas, 2008; Morante and Daelemans, 2009; Özgur and Radev, 2009). To help improve the information extraction tools, a corpus, called BioScope, has been annotated for speculation, negation and its linguistic scopes in biomedical texts (Szarvas et al., 2008).

However, when a biologist says he is interested in knowing all speculations about a biological entity (gene or protein for example) or a biological process, this claim concerns another meaning of the word "*speculation*". The latter is in this case more restrictive than previously, and close to the notion of hypothesis and uncertain proposal. This interest of biologists can be explained by different reasons. Firstly, since speculations give meaning to results, they sometimes carry more useful information than factual sentences. In addition, speculative sentences emphasize important data, which can be very useful in data-collection papers (genomic and post-genomic papers, see (Brent and Lok, 2005)). Finally, speculations can also give current trends or directions, by enabling the researchers to anticipate future experimental discoveries, or by suggesting other ways to envision biological problems and giving new ideas for future experiments (Blagosklonny and Pardee, 2002). Hence, despite its importance for biologists, the need to find speculation according to this view has been neglected until now in Natural Language

Proceedings of the Workshop on Negation and Speculation in Natural Language Processing, pages 32–40,
Uppsala, July 2010.

Processing. To our knowledge, the only work focusing specifically on this issue is the development of the rule-based system BioExcom (Desclés et al., 2009). Since BioExcom has obtained good results for detecting speculations but in a relatively small scale evaluation, it seems useful to test this tool on a large, unknown corpus like BioScope. Furthermore, it is important to compare in greater detail these two different approaches to characterize speculative sentences and see more precisely in what they differ.

We performed an automatic annotation of the BioScope corpus by BioExcom and we measured raw performance. We observed that the vast majority of the divergences between BioExcom results and BioScope were due to the criteria used for detecting a speculative sentence. We manually treated the diverging sentences in order to correctly evaluate BioExcom according to its own criteria.

The contributions of this paper are the following:

- We present an original approach for considering speculative sentences in bio text mining. It concerns the definition of a speculation, the criteria used to find it and its importance for biologists.
- We demonstrate the efficiency of BioExcom to recognize these statements on a large-scale evaluation with good results.
- According to this new approach we provide an annotated corpus freely available in order to be used by researchers.

2 Related work

Hyland (1995) has extensively studied hedging, from a linguistic perspective, (the term of hedging has been introduced by Lakoff (1972)) in biological papers. Hedging represents an absence of certainty and is employed to indicate either a lack of commitment to the truth value of an accompanying proposition; either a desire not to express that commitment categorically. Three main functions are outlined for hedging: weakening the strength of a statement, signalling uncertainty and expressing deference to the reader.

From a Natural Language Processing perspective, the first work was carried out by Light et al. (2004). After a linguistic study of the use of speculative language in MEDLINE abstracts, the authors tested the possibility of manually annotating the speculative sentences by experts and linguists whilst providing small annotation guidelines. They concluded that humans can reliably annotate speculative sentences but that it is not possible to distinguish between "high" speculations and "low" speculations. Furthermore they performed an experiment with different automated methods based principally on the retrieval of keywords. Wilbur et al. (2006) defined five qualitative dimensions for scientific text annotations. Two of them concerned speculative statements (certainty and evidence) and they defined various guidelines for annotating text using them.

Medlock and Briscone (2007) provided much more detailed guidelines for hedge detection. A linguist and a domain expert without any input into the guideline development process, labelled a publicly available dataset (FlyBase dataset, consisting of 6 papers on *Drosophila melanogaster*) in order to perform a probabilistic acquisition model and to test it. A separation of the acquisition and classification phases in semi-supervised machine learning was used.

Svarzas (2008) managed to devise a hedge classification in biomedical texts based on a weakly supervised selection of keywords. To evaluate their system, they manually annotated four papers of BMC Bioinformatics with the same criteria as Medlock and Briscone (2007). They obtained an F-Measure of 85% on the FlyBase dataset and an F-Measure of 75% on the BMC Bioinformatics dataset, when the training was carried out with the FlyBase dataset, demonstrating that the probability of their hedge classifiers is limited.

To recognize speculative sentences, Kilicoglu and Bergler (2008) also used speculative keywords from prior linguistic work and expanded them by WordNet, UMLS Specialist Lexicon, and by introducing syntactic patterns. In order to determine the speculative strength of sentences, they assigned speculative cues to weights in two methods. They worked on the two publicly available corpora, obtaining an F-Measure of 85% on the FlyBase data set and an F-Measure of 82% on the BMC Bioinformatics data.

The BioScope corpus is a manually annotated corpus for speculation and negation keywords (token level), and their linguistic scopes (sentence level) (Szarvas et al., 2008). This corpus is publicly available[1] and the annotation

[1] http://www.inf.u-szeged.hu/rgai/bioscope

process has been performed by two independent annotators and a chief linguist. In particular, the corpus consists of 9 full texts papers (five papers from the FlyBase dataset and four papers from the journal BMC Bioinformatics), and 1273 abstracts (from the Genia Corpus (Collier et al., 1999)). The annotation guidelines are provided and speculation is identified to uncertainty and hedging. However the criteria used here are not detailed very accurately, unlike the detailed work of Medlock and Briscone (2007).

In order to detect hedging but also their scope, two recent works were recently published. Morante and Daelemans (2009) present a machine learning system based on a previous system used to detect the scope of negation cues. Thus they show that the same scope finding system can be applied to both negation and hedging. They used the three subcropora of the BioScope corpus to test the efficiency of their system: the best F-Measures they obtained were 85% on the abstracts (10-fold cross-validation experiments), 72% on the Full Text Papers and 60% on the Clinical reports for hedge detection (training phase on the full abstract subcorpus). Özgur and Radev (2009) built also a system that detects speculations and their scopes in biomedical scientific texts. They performed a supervised classification task, where they classified the potential keywords as real speculation keywords or not by using a diverse set of linguistic features that represent the contexts of the keywords. They obtained F-Measure of 92% for the scientific abstracts (10-fold cross-validation experiments). The F-Measure for Full Text Papers (leave-one-out-cross-validation) was 83%.

3 BioExcom

EXCOM[2] (Djioua et al., 2006; Alrahabi, 2010) is a rule-based system using the computational principles of the Contextual Exploration processing (Desclés, 2006). EXCOM does not need any morpho-syntactic analysis and only requires one pre-treatment step of corpus segmentation into segments (which are sentences or sometimes clauses according to the punctuation) (Figure 1). EXCOM uses declarative rules built by linguists or domain experts and based on the search for linguistic markers in the text. The latter are hierarchically organized into the rules: they are either indicators (strong markers) or clues (complementary markers). Only the presence of indicators in the text triggers the associated CE rules, and then the additional clues can be searched for in the context defined by the rules, which is in our case the same sentence as the indicator (Figure 2).

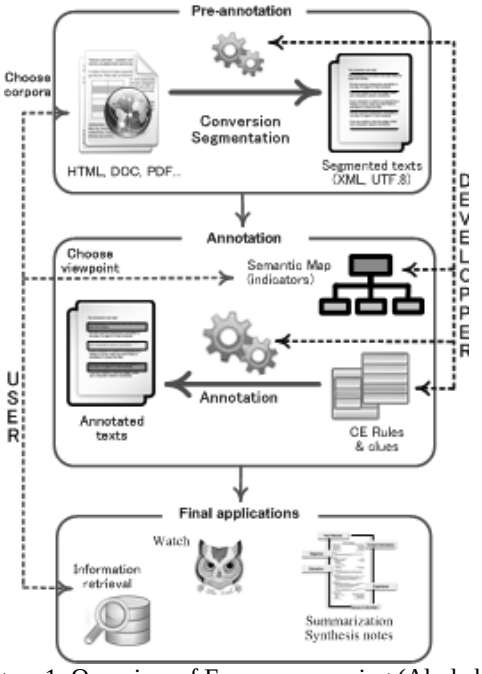

Figure 1: Overview of Excom processing (Alrahabi, 2010)

EXCOM allows the use of already annotated segments as markers, and also to order the rules and to use "negative" clues that cancel certain rules. The clues can be arranged between each other and versus the indicator (at its left, right or inside it). They are used to confirm, invalidate or specify an annotation carried by an indicator. If all the examined conditions of a rule are satisfied, EXCOM attributes the corresponding annotation to the designated segment. Generally the indicators are interconnected into graphs called "semantic maps" (Alrahabi and Desclés, 2008). EXCOM has been used for various tasks such as automatic summarization, relationships between concepts, categorization of bibliographic citations and reported speech.

BioExcom is the system that uses the annotation performed by EXCOM thanks to specific linguistic resources built for detecting speculations in biomedical papers (Desclés et al., 2009). Furthermore, BioExcom performs an indexation of the annotated segments in order to provide the possibility of searching for

[2] http://www.excom.fr/

specific terms. The rules used for the annotation processing are based on a precise semantic analysis of the multiple ways to express speculation performed by an expert in about seventy biological papers. BioExcom also categorizes speculative segments into "new speculation" (speculative sentences presented for the first time in the paper or not explicitly presented as prior speculation) and "prior speculation" (speculative sentences cited in the paper, but presented as having been proposed previously). BioExcom uses thirty rules, based on twelve indicator classes.

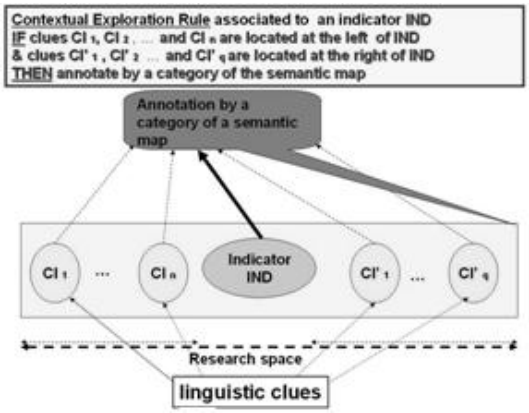

Figure 2: The contextual exploration principles: search for an indicator and then for some clues in a contextual space (a sentence or a clause in our case) according to some associated rules.

The criteria used to find speculations are described in detailed annotation guidelines, which are available on line[3]. Only the segments containing at least one linguistic element expressing speculation are considered. A speculation is defined as a proposal still not demonstrated about a biological issue and explicitly presented as uncertain in the paper. Thus, speculations are considered as potential sources of relevant information for biologists. Others types of statements such as deductions, conclusions, arguments and discussions are not considered as speculative but as intermediary statements, because they either present concepts more or less certain, or they do not make a proposal.

It is worth noting that contrary to other approaches, like Medlock (2008), the open questions are not speculative because they only ask a question about a biological problem without proposing a mechanism (sentence 1).

(1) *How endocytosis of DI leads to the activation of N remains to be elucidated.*

In order to better illustrate what a proposal is and how BioExcom performs annotation we can compare this sentence with the example (2) where BioExcom finds the indicator "*is not known*". Nevertheless, as the goal of this tool is not to extract all sentences expressing uncertainty, BioExcom searches for the clue "*whether*" at the right or at the left of the indicator. The presence of this clue indicates that, contrary to the sentence (1) with "*how*", the sentence (2) proposes a mechanism and consequently is a speculation.

(2) *Also, whether the signaling activity of Ser is similarly regulated by endocytosis is not known.*

The sentences which discuss a speculation without being informative about the content of the speculation are also not considered as speculations (sentence 3, extracted from the guidelines of Medlock (2008)). Indeed the goal of BioExcom is to detect a sentence explaining, at least partially, a speculation, and not to know, as the example (3), whether a speculation is purely speculative or is supported (or not) by facts. For the same reason, BioExcom extracts speculation without taking into account whether this speculation is denied or not.

(3) *The rescue of the D-mib mutant phenotype by ectopic expression of Neur strongly supports this interpretation*

At present, the evaluation process of BioExcom has only been performed on a relatively small scale (2 full-text papers read by five experts and containing 59 speculative sentences in total) and after the automatic annotation process (Desclés et al., 2009). Promising results have been reported for BioExcom in detecting speculations, by providing a high Precision of 98,3%, and a Recall of 95,1% (F-Measure of 96,7%). This rate is consistent with the goal of BioExcom to be a monitoring and decision support tool.

[3] http://www.bioexcom.net/

4 Raw evaluation

The Bioscope corpus (Szarvas et al., 2008) consists of three parts; namely medical free texts, biological full papers and biological scientific abstracts. However in this test, we only used two parts of the Bioscope corpus (full papers and abstracts) because we were preferentially interested in the biomedical scientific domain.

First, we cleaned the corpus from all element tags (angled brackets marking hedge keywords and parentheses marking scopes) and saved these documents into text files. The latter could then be automatically segmented into sentences or clauses and semantically annotated by BioExcom. As a result of this processing, BioExcom automatically extracted 341 segments from the Full Text Papers Corpus and 1489 segments from the Abstract Corpus (1830 segments in total)[4]. We could then compare our output files with the Bioscope Corpus which contained manual annotations. In this task we do not consider the categorization of BioExcom ("new speculation" and "prior speculation") and these annotated sentences are only considered as speculative. Thus, we obtained the results presented in Table 1. Consistent with the previous evaluation performed on BioExcom (Desclés et al., 2009), the Precision is high (approximately 93% in average, calculated from the total of segments of the two corpora). Nevertheless, the Recall dramatically falls to approximately 68% (in average) compared to the first evaluation (Recall of 93%).

	Precision	Recall	F-Measure
Full Text Papers	89,35	62,92	73,84
Abstracts	94,75	68,83	79,74

Table 1: Summary of raw results for BioExcom evaluation

Presented briefly are comments and some annotations performed by BioExcom which are in agreement with BioScope.

(4) *High conservation of residues near the inframe stop codon also suggests the importance of this region.*

(5) *Therefore, sets of genes identified from co-expression data may often contain multiple extraneous upstream sequences.*

(6) *To test the hypothesis that cortivazol acts in dex-resistant cells by making use of the residual GR found there, wild-type and dex-resistant clones were treated with various concentrations of cortivazol and the induction of GR mRNA was studied.*

(7) *To determine whether specific members of the NF-kappa B family contribute to this effect, we examined the abilities of different NF-kappa B subunits to act with Tat-I to stimulate transcription of HIV in Jurkat T-leukemia cells.*

If we compare the speculative keywords indicated in BioScope (underlined) and the markers used by BioExcom, we see some convergences ("*suggests*" and "*may*" for the sentences (4) and (5)), but also some divergences. In the sentence (6), BioExcom uses "*hypothesis that*" as an indicator in order to extract informative speculations and not only sentences containing the word "*hypothesis*". The example of the sentence (7) is much more illustrative for the differences: whereas only "*whether*" is taken into account as a keyword in BioScope, BioExcom uses "*to determine*" as an indicator and "*whether*" as a positive clue, allowing extracting only sentences containing a proposal (see example 2). The minimalist strategy (searching for the minimal unit that expresses speculation) followed by the annotators of BioScope for the keywords can explain these observations.

5 Corrected evaluation

Our goal was to evaluate the performance of BioExcom according to its own definition of speculation. To analyze the observed low Recall (Table 1), we assumed that all sentences presenting an agreement between both methods (manual annotation in BioScope and automatic annotation by BioExcom) were correctly annotated and we checked manually all the segments (984 segments) presenting a divergence of annotation. This checking was performed by a biologist as chief evaluator (writer

[4] The results of the annotation of BioScope corpus by BioExcom are publicly available: http://www.bioexcom.net/. It is worth mentioning that a few sentences were not segmented exactly in the same way by BioExcom and in the BioScope corpus (approximately 2% of divergences). We based all our calculations on BioExcom segmentation.

of the annotation guidelines) and two independent linguists, not allowed to communicate with each other. None of these evaluators knew the annotation performed by BioExcom or in BioScope (blind study). The conflicts were resolved by discussions during regular meetings and, in case of important uncertainty for at least two annotators, the sentences (54 in total) were not taken into account. The BioScope corpus re-annotated according to the criteria of BioExcom is publicly available[5].

We can divide the segments presenting a disagreement between BioExcom and BioScope into two groups depending on the annotation performed either automatically by BioExcom or manually in the BioScope corpus (Figure 3).

The first group consists of 865 segments which were considered as speculations in BioScope (representing around 36% of the total of annotated sentences in BioScope) and which were not annotated by BioExcom (see Figure 2). After manual checking, we found that only around one third of these segments in the corpus Full papers were speculative according to the criteria of BioExcom. This proportion was around one fourth in the BioScope corpus Abstract. The goal of BioExcom to avoid annotating open questions (see examples (1-2)) or lack of knowledge (sentence (8)) can explain some of the absences of annotation by BioExcom.

(8) *The exact role of the ubiquitination pathway in regulating apoptosis is still unclear.*

In other cases, some differences linked to the semantic conception of speculation play a role. Thus, the following sentences are considered as speculative in BioScope:

(9) *Furthermore, genetic and transplantation studies indicate that both Neur and Mib act in a non-autonomous manner [18,21,22,23,25,29], indicating that endocytosis of Dl is associated with increased Dl signalling activity.*

(10) *It can be deduced that the erythroid ALAS precursor protein has a molecular weight of 64.6 kd, and is similar in size to the previously isolated human housekeeping ALAS precursor of molecular weight 70.6 kd.*

Although these sentences correspond to hedging as it has been defined by Hyland (1995), we argue that the sentences (9-10) can be characterized more as a demonstration or a deduction with the expressions "*indicate that*" and "*it can be deduced*" than a speculation. According to the criteria used for establishing BioExcom rules, these sentences do not correspond to a speculation because they present things more or less as certain (see also (Thompson et al., 2008)). In this view, the case of "*indicate that*" is interesting to be detailed. Whereas many studies use it as a linguistic marker of speculation, Kilicoglu and Bergler (2008) recently moderated its speculative meaning by highlighting the additional need to take into account its context. The linguistic analysis used to establish the rules in BioExcom is much more restrictive and does not consider it as a marker.

It should also be mentioned that we noticed a few sentences which were incorrectly annotated in BioScope as speculation. Thus, the sentence (11) is rather, in our opinion, a definite statement ("*or*" can be replaced by "*and*").

(11) *Tandem copies of this 67-bp MnlI-AluI fragment, when fused to the chloramphenicol acetyltransferase gene driven by the conalbumin promoter, stimulated transcription in B cells but not in Jurkat T cells or HeLa cells.*

There are also some sentences annotated in Bioscope which are undoubtedly speculations but they were not identified by BioExcom. Thus, BioExcom has not annotated some speculative sentences because of the lack of some accurate markers into its linguistic resources. This is the case in the sentence (12): "*seems*" is recognized by BioExcom only when it is associated with some adjectives ("*seems probable*" for example). We can also cite the more ambiguous case of "*predicted*" (sentence (13)) which is also absent from the linguistic resources of BioExcom. Although a prediction can be considered as having a more particular status because it presents a low degree of uncertainty, we can consider that it should be taken into account because the predicted proposal remains hypothetical.

[5] http://www.bioexcom.net/

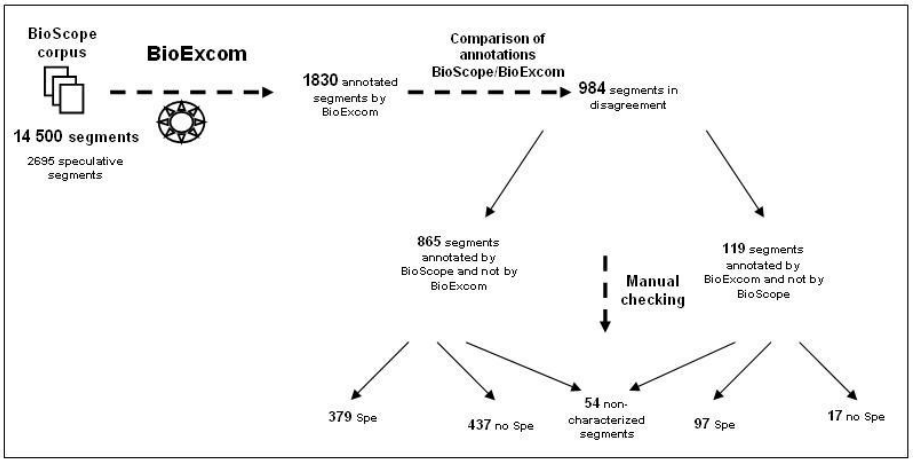

Figure 3: Results of the annotation of BioScope by BioExcom and of the manual checking (the calculations are based on BioExcom segmentation). Spe = speculative segment; no Spe = non speculative segment.

(12) *A transcription factor, NF-AT, which is essential for early T-cell gene activation, seems to be a specific target of cyclosporin A and FK506 action because transcription directed by this protein is blocked in T cells treated with these drugs, with little or no effect on other transcription factors such as AP-1 and NF-kappa B.*

(13) *Expression of full-length and mutant cDNA clones in bacteria reveal that the single HMG motif, which is predicted to contain two extended alpha-helical segments, is sufficient to direct the sequence-specific binding of TCF-1 alpha to DNA.*

Beside the lack of accurate markers, the absence of specific markers in some sentences does not allow the detection of some speculations by BioExcom. For example, in BioExcom, the ambiguity of the indicator "*could*" (past form or conditional form) is lifted by the presence of some positive clues expressing conditionality or possibility, such as "*if*" or "*whether*". But in the sentence (14), "*could*" has no clues to be disambiguated and although it is a speculation, BioExcom did not annotate it.

(14) *This method could be quite useful to detect not only CAG repeats in SBMA but also other polymorphic dinucleotide and trinucleotide repeats.*

The second group consists of segments annotated by BioExcom and not in BioScope (119 segments, see Figure 2). Around 80% of these sentences appeared to be speculative after manual examination. As an illustration, the following sentence is clearly a speculation ("*We hypothesize that*") but is not annotated in BioScope.

(15) *We hypothesize that a mutation of the hGR glucocorticoid-binding domain is the cause of cortisol resistance.*

Finally, based on these results, we decided to recalculate the Precision, Recall and F-Measure to be more in agreement with the notion of speculation as it has been described by Desclés et al. (2009) (last lines in Tables 2 and 3). Corrected Precision, Recall and F-Measure are respectively around 99%, 83% and 90% (averages calculated from the total of segments of the two corpora). These results are close to the first evaluation performed by Desclés et al. (2009), even if Recall is still lower than previously.

Obviously, our results are not directly comparable with the prior studies because BioExcom does not use exactly the same criteria to recognize the speculative sentences and consequently we re-annotated the BioScope corpus according to the criteria of BioExcom. One other difference is that the source used for linguistic resources and rules of BioExcom is completely different from the source of the corpus used for the evaluation, aside from the studies using directly BioScope like Morante and Daelemans (2009) or Özgur and Radev (2009). Nevertheless, considering that there are a few studies using a part of the BioScope cor-

pus, it can be interesting to mention that BioExcom achieves good performances in particular for Precision rate (Table 2 and 3).

		Precision	Recall	F-Measure
1	Fruit-fly dataset			85,08
1	BMC dataset			74,93
2	Fruit-fly dataset	85	86	85
2	BMC dataset	80	85	82
3	BioScope	75,35	68,18	71,59
4	BioScope	90,81	76,17	82,82
5	BioScope (corrected)	97,63	77,46	86,39

Table 2: Results reported in different publications and concerning the recognition of speculations in Scientific Full Text Papers, representing a part of the BioScope corpus: (1) (Szarvas, 2008), (2) (Kilicoglu and Bergler, 2008), (3) (Morante and Daelemans, 2009), (5) (Özgur and Radev, 2009), (5) BioExcom in this study

	Corpus	Precision	Recall	F-Measure
1	BioScope	90,81	79,84	84,77
2	BioScope	95,56	88,22	91,69
3	BioScope (corrected)	99,39	83,93	91,01

Table 3: Results reported in different publications and concerning the recognition of speculations in Scientific Abstracts, representing a part of the BioScope corpus: (1) (Morante and Daelemans, 2009) (2) (Özgur and Radev, 2009), (3) BioExcom in this study

6 Conclusion and Perspectives

Our aim was to test on a large scale (the manually annotated BioScope corpus) the rule based system BioExcom that automatically annotates speculations in biomedical papers. We observed an important disagreement between the two annotations and as a result, treated it manually. We put forward three principal reasons for the differences in annotation:

- The lack of certain linguistic markers in BioExcom (false negative): some of them have to be added in the system (for example "*seem*" or "*it appears*"). Some other markers are too ambiguous to be relevant (for example "*could*" without a positive clue).
- An error in the annotation of BioExcom or in BioScope (false positives): this is relatively rare, especially for BioExcom, which favors the Precision.
- The difference of criteria used to determine whether a sentence is speculative or not: this is the main reason and we discuss it hereafter.

Obtaining a good manual annotation is a recurrent problem in semantic annotation (Uren et al., 2006) and some studies have studied the disagreement in human judgment (Veronis, 1998). This phenomenon is undoubtedly found here since Szarvas et al. (2008) and Desclés et al. (2009) have reported the difficulty even for experts to detect speculations. But the disagreement between BioExcom and BioScope has to be seen almost as a conflict concerning the meaning given to the word "*speculation*" and if we agree on its definition, we demonstrated the efficiency of BioExcom.

Indeed, in order to provide good tools to biologists, computational linguists seek to extract only definite statements and consequently try to remove or to present separately hedging. Some of them have also tried to indicate the degree of uncertainty, in order to better characterize hedging. Despite the promising use of weighting of hedging cues (Kilicoglu and Bergler, 2008), the task of determining the strength of speculation appears to be difficult, even for biologists (Light et al., 2004). In another hand biologists can also consider speculation as a source of knowledge (for example, Thompson et al. (2008) categorize speculation as a type of knowledge), giving actual trends, new ideas and hypothesis useful for them. In this view, BioExcom extracts them according to more restrictive criteria and categorizes them into "prior" and "new" in order to better highlight speculations. Knowing the new speculations of a paper can reveal some of the real new output of it and so help to decide which paper is worth spending time. The categorization into prior speculation highlights the emergence of an idea which is taken into consideration by the scientific community and thus can also, at least partially, give an indication of its importance among the huge amount of speculations in the literature.

The availability of the corpora (raw BioExcom annotations and re-annotations of BioScope according to these criteria) could help to better take into account these views. In particular, these corpora will be useful to improve the rules of BioExcom for detecting speculations. And many of the other sentences belonging to hedging and discarded during the re-annotation process (previously annotated or not in BioScope) will serve to develop other semantic categories such as *demonstration*/*conclusion* or *deduction.*

Acknowledgments
We would like to thank J.P Desclés and M. Alrahabi for helpful discussions.

References

Alrahabi M (2010) EXCOM-2: a platform for automatic annotation of semantic categories: Conception, modeling and implementation. Application to the categorisation of quotations in French and Arabic, *PhD*, University of Paris-Sorbonne.

Alrahabi M, Desclés JP (2008) Automatic annotation of direct reported speech in Arabic and French, according to semantic map of enunciative modalities. *In 6th International Conference on Natural Language Processing, GoTAL*, Gothenburg, Sweden, pp 41-51

Blagosklonny MV, Pardee AB (2002) Conceptual biology: unearthing the gems. *Nature* 416: 373

Brent R, Lok L (2005) Cell biology. A fishing buddy for hypothesis generators. *Science* 308: 504-506

Collier, N., Park, H., Ogata, N., Tateishi, Y., Nobata, C., Ohta, T., Sekimizu, T., Imai, H., and Tsujii, J. (1999), "The GENIA project: corpus-based knowledge acquisition and information extraction from genome research papers", *Proc. Annual Meeting of the European Association for Computational Linguistics* (EACL-99), Bergen, Norway, June 8-12, pp.271-272.

Desclés JP, (2006) Contextual Exploration Processing for Discourse Automatic Annotations of Texts. *In FLAIRS 2006*, invited speaker, Melbourne, Florida, pp 281-284

Desclés J, Alrahabi M, Desclés JP (2009), BioExcom: Automatic Annotation and categorization of speculative sentences in biological literature by a Contextual Exploration processing, *In Proceedings of the 4th Language & Technology Conference* (LTC), Poznań, Poland

Djioua B, Flores JG, Blais A, Desclés JP, Guibert G, Jackiewicz A, Le Priol F, Nait-Baha L, Sauzay B (2006) EXCOM: an automatic annotation engine for semantic information. *In FLAIRS 2006*, Melbourne, Florida, 11-13 mai, Melbourne, Florida, pp 285-290

Hyland K (1995) The author in the text: Hedging Scientific Writing. *Hong Kong papers in linguistics and language teaching* 18: 33-42

Kilicoglu H, Bergler S (2008) Recognizing speculative language in biomedical research articles: a linguistically motivated perspective. *BMC Bioinformatics* 9 Suppl 11: S10

Lakoff G (1972) Hedges: a study in meaning criteria and the logic of fuzzy concepts. *In Chicago Linguistics Society Papers*, 8:183-228

Light M, Qiu XY, Srinivasan P (2004) The Language of Bioscience: Facts, Speculations, and Statements in Between. *In HLT-NAACL*, ed, *Workshop On Linking Biological Literature Ontologies And Databases*, pp 17-24

Medlock B, Briscone T (2007) Weakly Supervised Learning for Hedge Classification in Scientific Literature. *In AfC Linguistics, ed, Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, Prague, Czech Republic, pp 992-999

Morante R, Daelemans W (2009) Learning the scope of hedge cues in biomedical texts, *In Proceedings of the Workshop on BioNLP*, pp 28-36, Boulder, Colorado, USA, June 2009, ACL

Özgür A, Radev DR (2009) Detecting Speculations and their Scopes in Scientific Text, *In Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pp 1398-1407, Singapore, 6-7 August, 2009

Szarvas G (2008) Hedge classification in biomedical texts with a weakly supervised selection of keywords, *In Proceedings of ACL-08: HLT*, pp 281-289, Columbus, Ohio, USA, June 2008

Szarvas G, Vincze V, Farkas R, Csirik J (2008) The BioScope corpus: annotation for negation, uncertainty and their scope in biomedical texts. *In BioNLP ACL-2008* workshop

Thompson P, Venturi G, McNaught J, Montemagni S, Ananiadou S (2008) Categorising modality in biomedical texts. *Proceedings of the LREC 2008 Workshop on Building and Evaluating Resources for Biomedical Text Mining.* Marrakech, Morocco 2008.

Uren V, Cimiano P, Iria J, Handschuh S, Vargas-Vera M, Motta E, Ciravegna F (2006) Semantic annotation for knowledge management: Requirements and a survey of the state of the art. *In Web Semantics: Science, Services and Agents on the World Wide Web* Volume 4, Issue 1, January 2006, Pages 14-28

Véronis, J. (1998) A study of polysemy judgements and inter-annotator agreement, Programme and advanced papers of the *Senseval workshop* (pp. 2-4). Herstmonceux Castle (England)

Wilbur WJ, Rzhetsky A, Shatkay H (2006) New directions in biomedical text annotation: definitions, guidelines and corpus construction. *BMC Bioinformatics* 7: 35

Levels of Certainty in Knowledge-Intensive Corpora: An Initial Annotation Study

Aron Henriksson
DSV/KTH-Stockholm University
Sweden
aronhen@dsv.su.se

Sumithra Velupillai
DSV/KTH-Stockholm University
Sweden
sumithra@dsv.su.se

Abstract

In this initial annotation study, we suggest an appropriate approach for determining the level of certainty in text, including classification into multiple levels of certainty, types of statement and indicators of amplified certainty. A primary evaluation, based on pairwise inter-annotator agreement (IAA) using F_1-score, is performed on a small corpus comprising documents from the *World Bank*. While IAA results are low, the analysis will allow further refinement of the created guidelines.

1 Introduction

Despite ongoing efforts to codify knowledge, it is often communicated in an informal manner. In our choice of words and expressions, we implicitly or explicitly judge the certainty of the knowledge we wish to convey. This fact makes it possible to gauge the reliability of knowledge based on the subjective perspective of the author.

As knowledge is often difficult to ascertain, it seems reasonable to regard knowledge on a continuum of varying degrees of certainty, as opposed to a binary (mis)conception. This corresponds to the notion of *epistemic modality*: the degree of confidence in, or commitment to, the truth of propositions (Hyland, 1998). *Hedging* is a means of affecting *epistemic modality* by qualifying propositions, realized through tentative words and expressions such as *possibly* and *tends to*.

A holistic perspective on certainty—in which not only speculation is considered, but also signs of increased certainty—requires a classification into various levels. Applying such an approach to knowledge-intensive corpora, it may in due course be possible to evaluate unstructured, informal knowledge. This would not least be valuable to organizational knowledge management practices, where it could provide a rough indicator of reliability in internal *knowledge audits*.

2 Related Research

The *hedging* concept was first introduced by Lakoff (1973) but has only really come into the spotlight in more recent years. Studies have mainly taken place in the biomedical domain, with Hyland's (1998) influential work investigating the phenomenon in scientific research articles. Speculative keywords and negations, along with their linguistic scopes, are annotated in the *BioScope* corpus by Vincze et al. (2008), which contains a large collection of medical and biological text (scientific articles and abstracts, as well as radiology reports). After several iterations of refining their guidelines, they report IAA values ranging from 77.6 to 92.37 F_1-score for speculative keywords (62.5 and 95.54 F_1-score for full scope). This corpus is freely available and has been used for training and evaluation of automatic classifiers, see e.g. Morante and Daelemans (2009). One of the main findings is that hedge cues are highly domain-dependent. Automatic identification of other private states, including opinions, represents a similar task, see e.g. Wiebe et al. (2005). Diab et al. (2009) study annotation of committed and non-committed belief and show that automatic tagging of such classes is feasible. A different annotation approach is proposed by Rubin et al. (2006), in which certainty in newspaper articles is categorized along four dimensions: *level*, *perspective*, *focus* and *time*. Similarly, five dimensions are used in Wilbur et al. (2006) for the creation of an annotated corpus of biomedical text: *focus*, *polarity*, *certainty*, *evidence* and *directionality*.

3 Method

Based on previous approaches and an extensive literature review, we propose a set of guidelines that

Proceedings of the Workshop on Negation and Speculation in Natural Language Processing, pages 41–45,
Uppsala, July 2010.

(1) incorporates some new features and (2) shifts the perspective to suit knowledge-intensive corpora, e.g. comprising organizational knowledge documents. Besides categorization into levels of certainty, this approach distinguishes between two types of statement and underscores the need to take into account words and expressions that add certainty to a proposition.

A small corpus of 10 *World Bank* documents—a publicly available resource known as *Viewpoints* (The World Bank Group, 2010)—is subsequently annotated in two sets by different annotators. The corpus is from a slightly different domain to those previously targeted and represents an adequate alternative to knowledge documents internal to an organization by fulfilling the criterion of knowledge intensity. The process is carried out in a *Protégé* plugin: *Knowtator* (Ogren, 2006). Pairwise IAA, measured as F_1-score, is calculated to evaluate the feasibility of the approach.

Statements are annotated at the clause level, as sentences often contain subparts subject to different levels of certainty. These are not predefined and the span of classes is determined by the annotator. Furthermore, a distinction is made between different types of statement: statements that give an *account* of something, typically a report of past events, and statements that express concrete knowledge *claims*. The rationale behind this distinction is that text comprises statements that make more or less claims of constituting knowledge. Thus, knowledge *claims*—often less prevalent than *accounts*—should be given more weight in the overall assessment, as the application lies in automatically evaluating the reliability of informal knowledge. Assuming the view of knowledge and certainty as continuous, it is necessary to discretize that into a number of intervals, albeit more than two. Hence, *accounts* and *claims* are categorized according to four levels of certainty: *very certain*, *quite certain*, *quite uncertain* and *very uncertain*. In addition to the statement classes, four indicators make up the total of twelve. We introduce *certainty amplifiers*, which have received little attention in previous work. These are linguistic features that add certainty to a statement, e.g. words like *definitely* and expressions like *without a shadow of a doubt*. *Hedging indicators*, on the other hand, have gained much attention recently and signify uncertainty. The *source hedge* class is applicable to instances where the source of *epistemic judgement* is stated explicitly, yet only when it provides a hedging function (e.g. *some say*). *Modality strengtheners* are features that strengthen the effect of *epistemic modality* when used in conjunction with other (un)certainty indicators—but alone do not signify any polarity orientation—and may be in the form of vagueness (e.g. <could be> *around* that number) or quantity gradations (e.g. *very* <sure>).

4 Results

The corpus contains a total of 772 sentences, which are annotated twice: set #1 by one annotator and set #2 by five annotators, annotating two documents each. The statistics in Table 1 show a discrepancy over the two sets in the number of classified statements, which is likely due to difficulties in determining the scope of clauses. There are likewise significant differences in the proportion between *accounts* and *claims*, as had been anticipated.

Accounts		**Claims**	
Set #1	Set #2	Set #1	Set #2
726	574	395	393

Table 1: Frequencies of accounts and claims.

Despite the problem of discriminating between *accounts* and *claims*, they seem to be susceptible to varying levels of certainty. The average distribution of certainty for *account* statements is depicted in Figure 1. As expected, an overwhelming majority (87%) of such statements are *quite certain*, merely relating past events and established facts.

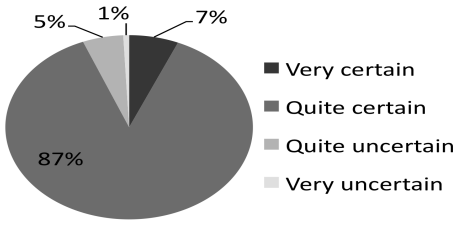

Figure 1: Average distribution of certainty in *account* statements.

By comparison, knowledge *claims* are more commonly hedged (23%), although the majority is still *quite certain*. Interestingly, *claims* are also expressed with added confidence more often than *accounts*—around one in every ten *claims*.

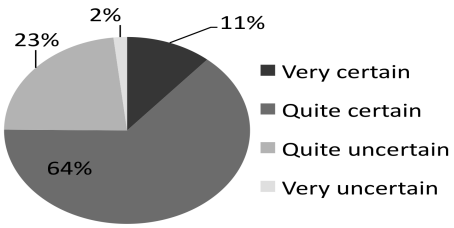

Figure 2: Average distribution of certainty in knowledge *claims*.

As expected, the most common indicator is of hedging. Common cues include *may*, *can*, *might*, *could*, *indicate(s)*, *generally* and *typically*. Many of these cues are also among the most common in the biomedical sub-corpus of *BioScope* (Vincze et al., 2008). It is interesting to note the fairly common phenomenon of *certainty amplifiers*. These are especially interesting, as they have not been studied much before, although Wiebe et al. (2005) incorporate *intensity ratings* in their annotation scheme. There is agreement on words like *clearly*, *strongly* and *especially*.

Indicator	**Set #1**	**Set #2**
Certainty amplifier	61	29
Hedging indicator	151	133
Source hedge	0	40
Modality strengthener	9	122

Table 2: Frequency of indicators

To evaluate the approach, we calculate IAA by pairwise F_1-score, considering set #1 as the gold standard, i.e. as correctly classified, in relation to which the other subsets are evaluated. We do this for exact matches and partial matches[1]. For exact matches in a single document, the F_1-score values range from an extremely low 0.09 to a somewhat higher—although still poor—0.52, yielding an overall average of 0.28. These results clearly reflect the difficulty of the task, although one has to keep in mind the impact of the discrepancy in the number of annotations. This is partly reflected in the higher overall average for partial matches: 0.41.

Certainty amplifiers and *hedging indicators* have F_1-scores that range up to 0.53 and 0.55 respectively (ditto for partial matches) in a single document. Over the entire corpus, however, the averages come down to 0.27 for *certainty amplifiers* (0.30 for partial matches) and 0.33 for *hedging indicators* (0.35 for partial matches).

Given the poor results, we want to find out whether the main difficulty is presented by having to judge certainty according to four levels of certainty, or whether it lies in having to distinguish between types of statement. We therefore generalize the eight statement-related classes into a single division between *accounts* and *claims*. Naturally, the agreement is higher than for any single class, with 0.44 for the former and 0.41 for the latter. A substantial increase is seen in partial matches, with 0.70 for *accounts* and 0.55 for *claims*. The results are, however, sufficiently low to conclude that there were real difficulties in distinguishing between the two.

Statement Type	**Exact F_1**	**Partial F_1**
Account	0.44	0.70
Claim	0.41	0.55

Table 3: Pairwise IAA per statement type, F_1-scores for exact and partial matches.

We subsequently generalize the eight classes into four, according to their level of certainty alone. The results are again low: *quite certain* yields the highest agreement at 0.47 (0.76 for partial matches), followed by *quite uncertain* at 0.24 (0.35 for partial matches). These numbers suggest that this part of the task is likewise difficult. The rise in F_1-scores for partial matches is noteworthy, as it highlights the problem of different interpretations of clause spans.

Certainty Level	**Exact F_1**	**Partial F_1**
Very certain	0.15	0.15
Quite certain	0.47	0.76
Quite uncertain	0.24	0.35
Very uncertain	0.08	0.08

Table 4: Pairwise IAA per certainty level, F_1-scores for exact and partial matches

5 Discussion

In the guidelines, it is suggested that the level of certainty can typically be gauged by identifying the number of indicators. There is, however, a serious drawback to this approach. Hedging indicators, in particular, are inherently uncertain to different degrees. Consider the words *possibly* and

[1]Partial matches are calculated on a character level while exact matches are calculated on a token level.

probably. According to the guidelines, a single occurrence of either of these hedging indicators would normally render a statement *quite uncertain*. Giving freer hands to the annotator might be a way to evade this problem; however, it is not likely to lead to any more consistent annotations. Kilicoglu and Bergler (2008) address this by assigning weights to hedging cues.

A constantly recurring bone of contention is presented by the relationship between certainty and precision. One of the hardest judgements to make is whether imprecision, or vagueness, is a sign of uncertainty. Consider the following example from the corpus:

> *Cape Verde had virtually no private sector.*

Clearly, this statement would be more certain if it had said: *Cape Verde had no private sector.* However, *virtually no* could be substituted with, say, *a very small*, in which case the statement would surely not be deemed uncertain. Perhaps precision is a dimension of knowledge that should be analyzed in conjunction with certainty, but be annotated separately.

6 Conclusion

There are, of course, a number of ways one can go about annotating the level of certainty from a knowledge perspective. Some modifications to the approach described here are essential—which the low IAA values are testament to—while others may be worth exploring. Below is a selection of five key changes to the approach that may lead to improved results:

1. *Explicate statement types*. Although there seems to be a useful difference between the two types, the distinction needs to be further explicated in the guidelines.

2. *Focus on indicators*. It is clear that indicators cannot be judged in an identical fashion only because they have been identified as signifying either certainty or uncertainty. It is not simply the number of occurrences of indicators that determines the level of certainty but rather how *strong* those indicators are. A possible solution is to classify indicators according to the level of certainty they affect.

3. *Discard rare classes*. Very rare phenomena that do not have a significant impact on the overall assessment can be sacrificed without affecting the results negatively, which may also make the task a little less difficult.

4. *Clarify guidelines*. A more general remedy is to clarify further the guidelines, including instructions on how to determine the scope of clauses; alternatively, predefine them.

5. *Instruct annotators*. Exposing annotators to the task would surely result in increased agreement, in particular if they agree beforehand on the distinctions described in the guidelines. At the same time, you do not want to steer the process too much. Perhaps the task is inherently difficult to define in detail. Studies on how to exploit subjective annotations might be interesting to explore, see e.g. Reidsma and op den Akker (2008).

In the attempt to gauge the reliability of knowledge, incorporating multiple levels of certainty becomes necessary, as does indicators of increased certainty. Given the similar rates of agreement on *hedging indicators* and *certainty amplifiers* (0.33 and 0.27 respectively; 0.30 and 0.35 for partial matches), the latter class seem to be confirmed. It is an existing and important phenomenon, although—like *hedging indicators*—difficult to judge. Moreover, a differentiation between types of statement is important due to their—to different degrees—varying claims of constituting knowledge. An automatic classifier built on such an approach could be employed with significant benefit to organizations actively managing their collective knowledge. The advantage of being aware of the reliability of knowledge are conceivably manifold: it could, for instance, be (1) provided as an attribute to end-users browsing documents, (2) used as metadata by search engines, (3) used in *knowledge audits* and *knowledge gap analyses*, enabling organizations to learn when knowledge in a particular area needs to be consolidated. It is, of course, also applicable in a more general information extraction sense: information that is extracted from text needs to have a certainty indicator attached to it.

A dimension other than certainty that has a clear impact on knowledge is precision. It would be interesting to evaluate the reliability of knowledge based on a combination of certainty and precision.

The annotated *World Bank* corpus will be made available for further research on the Web.

References

Mona T. Diab, Lori Levin, Teruko Mitamura, Owen Rambow, Vinodkumar Prabhakaram, and Weiwei Guo. 2009. Committed belief annotation and tagging. In *Proceedings of the Third Linguistic Annotation Workshop, ACL-IJCNLP*, pages 68–73, Suntec, Singapore, August. ACL and AFNLP.

Ken Hyland. 1998. *Hedging in Scientific Research Articles*. John Benjamins Publishing Company, Amsterdam/Philadelphia.

Halil Kilicoglu and Sabine Bergler. 2008. Recognizing speculative language in biomedical research articles: a linguistically motivated perspective. *BMC Bioinformatics*, 9.

George Lakoff. 1973. Hedges: A study in meaning criteria and the logic of fuzzy concepts. *Journal of Philosophical Logic*, 2:458–508.

Roser Morante and Walter Daelemans. 2009. Learning the scope of hedge cues in biomedical texts. In *Proceedings of the Workshop on BioNLP*, pages 28–36, Boulder, Colorado, June. Association for Computational Linguistics.

Philip V. Ogren. 2006. Knowtator: a protégé plug-in for annotated corpus construction. In *Proceedings of the 2006 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, pages 273–275, Morristown, NJ, USA. Association for Computational Linguistics.

Dennis Reidsma and Rieks op den Akker. 2008. Exploiting 'subjective' annotations. In *HumanJudge '08: Proceedings of the Workshop on Human Judgements in Computational Linguistics*, pages 8–16, Morristown, NJ, USA. Association for Computational Linguistics.

Victoria L. Rubin, Elizabeth D. Liddy, and Noriko Kando. 2006. Certainty identification in texts: Categorization model and manual tagging results. In *Computing Affect and Attitutde in Text: Theory and Applications*. Springer.

The World Bank Group. 2010. Documents & Reports. `http://go.worldbank.org/3BU2Z3YZ40`, *Accessed May 13, 2010*.

Veronika Vincze, György Szaarvas, Richárd Farkas, György Móra, and János Csirik. 2008. The bioscope corpus: biomedical texts annotated for uncertainty, negation and their scopes. *BMC Bioinformatics*, 9.

Janyce Wiebe, Theresa Wilson, and Claire Cardie. 2005. Annotating expressions of opinions and emotions in language. *Language Resources and Evaluation*, 39:165–210.

J. W. Wilbur, A. Rzhetsky, and H. Shatkay. 2006. New directions in biomedical text annotation: definitions, guidelines and corpus construction. *BMC Bioinformatics*, 7:356+, July.

Importance of negations and experimental qualifiers in biomedical literature

Martin Krallinger
Struct. Biol. and Biocomp. Prog.
Spanish National Cancer Center, Madrid, Spain.
mkrallinger@cnio.es

Abstract

A general characteristic of most biomedical disciplines is their primarily experimental character. Discoveries are obtained through molecular biology and biochemical techniques that allow understanding of biological processes at the molecular level. To qualify biological events, it is of practical significance to detect specific types of negations that can imply either that a given event is not observed under specific conditions or even the opposite, that a given event is true by altering the bio-entities studied (e.g. introducing specific modifications like mutations). Of special interest is also to determine if a detected assertion is linked to experimental support provided by the authors. Finding experimental qualifier cues and detecting experimental technique mentions is of great interest to the biological community in general and particularly for annotation databases. A short overview of different types of negations and biological qualifiers of practical relevance will be provided.

1 Biological Annotations

In line with the rapid accumulation of biological literature and the growing number of large-scale experiments in biomedicine, it is becoming more important to capture essential facts contained in the literature and storing them in form of biological annotations. Such annotations usually consist in structured database records, where biological entities of relevance, like genes or proteins are associated to controlled vocabularies that are useful to describe the most relevant aspects of these entities (their function, localization, processes or pathways they participate in or implications in diseases). Also specific types of relations between bio-entities (e.g. physical or regulatory interactions) are manually extracted from the literature. For biological interpretation and to determine the reliability of annotations it is crucial to capture both negative annotations, whether a given relation has been studied experimentally and does not occur, as well as to determine the experimental method used to study the bio-entity of interest. For instance, the value of in vitro generated results, or those obtained by large-scale experiments have a different significance compared to those generated in vivo. The most relevant biological annotations contained in databases and constructed manually by expert curators are linked to experimental qualifiers. Such experimental qualifiers can range from simple method terms to more sophisticated ontologies or hierarchical terminologies. Experimental qualifiers used to annotate biological entities are for instance provided by the Proteomics Standards Initiative Molecular Interaction (PSI-MI) ontology, (Orchard S, Kerrien S., 2010) the Evidence Codes of Gene Ontology (GO) (Rogers MF, Ben-Hur A, 2010) or the Open REGulatory ANNOtation (ORegAnno) database Evidence Types.

2 Importance of Negations in Biomedicine

There is an increasing interest to extract from the literature negative associations. For instance, one of the most popular biological annotation efforts, Gene Ontology Annotation (GOA), also supports the annotation of $'NOT'$ relations (association.is_not) to be able to represent these types of relations in their annotation data. In GO, such relations are labeled using $'NOT'$ in the qualifier column for a particular annotation. This negation qualifier is applied to provide an explicit note that the bio-entity is not associated with a given GO term. This is important when a GO term might otherwise be expected to apply to a bio-entity, but an experiment proves otherwise. Negative asso-

Proceedings of the Workshop on Negation and Speculation in Natural Language Processing, pages 46–49,
Uppsala, July 2010.

ciations are also used when a cited reference explicitly states a negation event, e.g. in the form of: bio-entity X is not found in the location Y. In addition to annotation efforts there are a range of scenarios where extraction of negative events are of practical importance, these are described in the following subsections.

2.1 Negations and Negative Controls

A common setting in experimental biology is to use controls to avoid alternative explanations of results and to minimize experimental artifacts. Negative controls corroborate that the experimental outcome is not due to some sort of unrelated effect; it serves to minimize false positives and can serve as a background observation. The underlying assumption of negative controls is that one assumes in advance that the result should be negative, i.e. no significant effect should be obtained. Such negative controls are mainly expressed in the literature using negations. For instance in case of protein-protein interaction experiments, a negative control could be to demonstrate that a signal is only obtained when the two interactor proteins are present, and not when the label (tag-protein) alone is given to each interactor individually. To illustrate this aspect consider the example sentences provided below:

- *Our results show that, when AGG1 is present in the matrix, it shows a strong ability to bind 35S-labeled AGB1, whereas GST alone is not able to bind any detectable AGB1.*
- *GST alone did not interact with FKHR even in the presence of E2 (Fig. 2B, lane 5), indicating the specific interaction between ER and FKHR.*
- *35S-labeled in vitrotranslated FBXO11 bound to immobilized GST-p53 (lane 3) but not GST alone (lane 2).*
- *PKC bound to GST-RINCK1 (lane 2) but not to GST alone (lane 1), revealing that PKC binds to RINCK directly.*

In those example cases, GST (alone) would represent the negative control. Only in presence of the interactor proteins a signal should be observed, if GST alone is present the assumption is that no signal should be obtained. Negative controls are crucial for interpretation of the actual experimental outcome.

2.2 Negative associations in medical and population genetics

A considerable effort is being made to detect genes and mutations in genes that have implications in the susceptibility of complex disorders. Naturally occurring variations in the sequence of genes, often called polymorphisms might have a deleterious, protective or no associations at all to a pathologic condition. Not only to capture deleterious and protective mutations, but also those that do not have any effect is important to aid in the interpretation of mutations observed in patients. This is especially true taking into account the increasing use of molecular screening technologies and personalized medicine in the clinical domain. Example cases of negative associations between genes and mutations to disease conditions derived from PubMed abstracts can be seen below:

- *CC16 gene may be not a susceptibility gene of asthmatic patients of Han population in southwest China.*
- *The FZD3 gene might not play a role in conferring susceptibility to major psychosis in our sample.*
- *Apolipoprotein E gene polymorphism is not a strong risk factor for diabetic nephropathy and retinopathy in Type I diabetes: case-control study.*
- *In view of this evidence, it is likely that the SIGMAR1 gene does not confer susceptibility to schizophrenia.*
- *Thus, this SNP in the PGIS gene is not associated with EH.*
- *The gene encoding GABBR1 is not associated with childhood absence epilepsy in the Chinese Han population.*
- *We did not find an association between OCD, family history for OCD, and the COMT gene polymorphism.*

Such negative associations can be useful for the interpretation of relevance of genes for certain conditions, enabling filtering un-relevant genes and improving target selection for more detailed molecular examinations.

2.3 Toxicology and negations

A simplified view of toxicology experiments is to distinguish, given the administration of different amounts of a specific compound or drug (e.g. low, medium and high dosage) during predefined time spans, between toxic and non-toxic effects. Such effects can be examined in animal models like rats or mini-pigs by examining a series of aspects, such as hematological parameters, organ histological properties (tissue alterations and size of organs), biochemical parameters, and changes in food/water consumption or fertility. Usually animals to which specific amounts of the compound has been administered are compared to control cases. Here it is important to determine also three kinds of negative associations: (1) under which conditions a given parameter or tissue has not been negatively affected (save dosage, non-toxic), (2) which compound did not show the desired beneficial effect (e.g. was not effective in treating the pathologic condition) and (3) under which administration conditions a compound was not save. Example sentences illustrating these negative associations are:

- *Morphological evaluation showed that 1-BP did not cause morphological changes in seminiferous epithelium, but 2-BP treatment resulted in the disappearance of spermatogonia, atrophy of the seminiferous tubules and degeneration of germ cells..*
- *This is an indication that the extracts may not be completely safe in male rats when continuously administered for 14days.*
- *Histopathologic analysis of the vital organs revealed no significant lesions in the brain, liver, kidney, heart, spleen, ovary, and testis.*
- *The extract did not produce any significant (P>0.05) changes in the mean concentrations of urea, creatinine, Na+, K+, and Cl-ions of rats in the extract treated groups compared to that of control.*

2.4 Experimentally altered bio-entities and negations

In order to characterize certain biological associations, it is a common practice to alter the bio-entity of interest, with the assumption that a given observation should change upon alteration. This is the case of mutations or deletions experimentally introduced to gene or protein sequences, with the underlying assumption that the mutated or truncated protein/gene should loose it ability to bind or regulate another bio-entity, or even be non-functional. Such mutations are useful to pin down the actual biologically relevant functional parts of bio-entities, which are usually of great therapeutic importance (as target sites to inhibit certain bio-entities or interactions). Such cases can be seen in the example sentences provided below:

- *Accordingly, this p73 N-terminal deletion was unable to activate transcription or to induce apoptosis.*
- *The G62D mutant did not bind AMP at all.*
- *The resulting mutant SOS3 protein was not able to interact with the SOS2 protein kinase and was less capable of activating it.*
- *MYB4 did not localize to the nucleus in the sad2 mutant, suggesting that SAD2 is required for MYB4 nuclear trafficking.*

In these example cases, altered bio-entities did not display the biological function of their wild type (unaltered) counterparts.

3 Experimental qualifiers

Biological annotation efforts are primarily concerned about experimentally confirmed events. Despite the importance of experimental qualifiers, only limited effort has been made to construct comprehensive resources to retrieve assertions that have experimental support and to construct useful lexical resources and thesauri of experimental evidence techniques. To detect novel protein interactions that have been experimentally characterized in the biomedical literature was one of the tasks posed in the BioCreative challenge, a community effort to assess text-mining tools developed for the biomedical domain (Krallinger M, et al, 2008). Also some systems to detect technical term mentions have been developed such as Termine. A range of recurrent cues relevant for experimental qualifiers can be observed in the literature, some of the most relevant ones are summarized in the table 1.

Using such experimental evidence cues together with linguistic patterns and NLP techniques it is feasible to determine whether a given event described in the literature has some sort of experi-

Cue	Pattern	PMID
reveal	METHOD revealed that EVENT	12506203
show	METHOD showed that EVENT	17189287
demonstrate	METHOD demonstrated that EVENT	18466309
study	EVENT was studied by METHOD	15147239
identify	EVENT identified in METHOD	10905349
prove	EVENT proved by METHOD	16354655
analyze	EVENT analyzed by METHOD	9477575
determine	EVENT determined by METHOD	12006647
confirm	EVENT confirmed using METHOD	10788494
obtain	EVENT obtained by METHOD	16582012
support	EVENT supported by METHOD	18156215
corroborate	EVENT corroborated using METHOD	15757661
validate	EVENT validated by METHOD	17287294
verify	EVENT verified by METHOD	18296724
detect	EVENT detected with METHOD	14581623
discover	EVENT discovered by METHOD	11251078
observe	EVENT observed using METHOD	16778013
test	EVENTwas tested using METHOD	14646219

Table 1: Experimental evidence cue terms.

mental qualifier associated to it. The simplest patterns of this sort would be for instance:

- *METHOD cue (a|that|novel|the|this)*
- *METHOD cue that*
- *as cue by METHOD*
- *was cue by METHOD*
- *cue (in|by|here by|using|via|with) METHOD*

Applying such patterns can be useful to construct automatically an experimental technique dictionary that can be handcrafted to enrich existing evidential qualifier resources. Nevertheless, linking automatically extracted experiment terms to controlled vocabularies used for annotation in biology is still a challenging task that need more manually labeled textual data. Some example sentences illustrating the usefulness of experimental evidence cues can be seen below:

- *Gel-shift and co-immunoprecipitation assays have revealed that GT-1 can interact with and stabilize the TFIIA-TBP-TATA complex.*
- *By yeast two-hybrid assays, we demonstrate an interaction of APC2 with two other APC/C subunits.*
- *The specificity of interaction of VIRP1 with viroid RNA was studied by different methodologies, which included Northwestern blotting, plaque lift, and electrophoretic mobility shift assays.*
- *A complex containing Mus81p and Rad54p was identified in immunoprecipitation experiments.*
- *In addition, we proved by affinity chromatography that NaTrxh specifically interacts with S-RNase.*

Acknowledgments

I would like to thank Yasmin Alam-Farugue (GOA team at EBI) for useful information on the annotation of negative associations in GOA and Roser Morante for important feedback and suggestions on this topic.

References

MF. Rogers and A. Ben-Hur. 2010. The use of gene ontology evidence codes in preventing classifier assessment bias., *Bioinformatics*, 25(9):1173-1177.

M. Krallinger and F. Leitner and C. Rodriguez-Penagos and A. Valencia 2008. Overview of the protein-protein interaction annotation extraction task of BioCreative II., *Genome Biol.*, Suppl 2:S1.

S. Orchard and S. Kerrien 2010. Molecular interactions and data standardisation., *Methods Mol Biol.*, 604:309-318

Negation and Modality in Distributional Semantics

Ed Hovy
Information Sciences Institute
University of Southern California
hovy@isi.edu

Abstract

In Natural Language Processing, negation and modality have mostly been handled using the older, pre-statistical methodologies of formal representations subject to rule-based processing. This fits the traditional treatment of negation and modality in logic-based knowledge representation and linguistics. However, in modern-day statistics-based NLP, how exactly negation and modality should be taken into account, and what role these phenomena play overall, is much less clear. The closest statistics-based NLP gets to semantics at this time is lexical-based word distributions (such as used in word sense disambiguation) and topic models (such as produced by Latent Dirichlet Allocation). What exactly in such representations should a negation or a modality actually apply to? What would, or should, the resulting effects be? The traditional approaches are of little or no help.

In this talk I argue that neither model is adequate, and that one needs a different model of semantics to be able to accommodate negation and modality. The traditional formalisms are impoverished in their absence of an explicit representation of the denotations of each symbol, and the statistics-based word distributions do not support the compositionality required of semantics since it is unclear how to link together two separate word distributions in a semantically meaningful way. A kind of hybrid, which one could call Distributional Semantics, should be formulated to include the necessary aspects of both: the ability to carry explicit word associations that are still partitioned so as to allow negation and modality to affect the representations in intuitively plausible ways is what is required.

I present a specific model of Distributional Semantics that, although still rudimentary, exhibits some of the desired features. I explore the possibilities for accommodating the phenomena of negation and modality. The talk poses many more questions than it answers and is an invitation to consider Distributional Semantics as a model for richer and more semantics-oriented statistics-based NLP.

Proceedings of the Workshop on Negation and Speculation in Natural Language Processing, page 50,
Uppsala, July 2010.

What's Great and What's Not: Learning to Classify the Scope of Negation for Improved Sentiment Analysis

Isaac G. Councill
Google, Inc.
76 Ninth Avenue
New York, NY 10011
icouncill@google.com

Ryan McDonald
Google, Inc.
76 Ninth Avenue
New York, NY 10011
ryanmcd@google.com

Leonid Velikovich
Google, Inc.
76 Ninth Avenue
New York, NY 10011
leonidv@google.com

Abstract

Automatic detection of linguistic negation in free text is a critical need for many text processing applications, including sentiment analysis. This paper presents a negation detection system based on a conditional random field modeled using features from an English dependency parser. The scope of negation detection is limited to explicit rather than implied negations within single sentences. A new negation corpus is presented that was constructed for the domain of English product reviews obtained from the open web, and the proposed negation extraction system is evaluated against the reviews corpus as well as the standard BioScope negation corpus, achieving 80.0% and 75.5% F1 scores, respectively. The impact of accurate negation detection on a state-of-the-art sentiment analysis system is also reported.

1 Introduction

The automatic detection of the scope of linguistic negation is a problem encountered in wide variety of document understanding tasks, including but not limited to medical data mining, general fact or relation extraction, question answering, and sentiment analysis. This paper describes an approach to negation scope detection in the context of sentiment analysis, particularly with respect to sentiment expressed in online reviews. The canonical need for proper negation detection in sentiment analysis can be expressed as the fundamental difference in semantics inherent in the phrases, "this is great," versus, "this is not great." Unfortunately, expressions of negation are not always so syntactically simple.

Linguistic negation is a complex topic: there are many forms of negation, ranging from the use of explicit cues such as "no" or "not" to much more subtle linguistic patterns. At the highest structural level, negations may occur in two forms (Givón, 1993): morphological negations, where word roots are modified with a negating prefix (e.g., "dis-", "non-", or "un-") or suffix (e.g., "-less"), and syntactic negation, where clauses are negated using explicitly negating words or other syntactic patterns that imply negative semantics. For the purposes of negation scope detection, only syntactic negations are of interest, since the scope of any morphological negation is restricted to an individual word. Morphological negations are very important when constructing lexicons, which is a separate but related research topic.

Tottie (1991) presents a comprehensive taxonomy of clausal English negations, where each form represents unique challenges for a negation scope detection system. The top-level negation categories – denials, rejections, imperatives, questions, supports, and repetitions – can be described as follows:

- **Denials** are the most common form and are typically unambiguous negations of a particular clause, such as, "There is no question that the service at this restaurant is excellent," or, "The audio system on this television is not very good, but the picture is amazing."

- **Rejections** often occur in discourse, where one participant rejects an offer or suggestion of another, e.g., "Can I get you anything else? No." However, rejections may appear in expository text where a writer explicitly rejects a previous supposition or expectation, for instance, "Given the poor reputation of the manufacturer, I expected to be disappointed with the device. This was not the case."

- **Imperatives** involve directing an audience

Proceedings of the Workshop on Negation and Speculation in Natural Language Processing, pages 51–59,
Uppsala, July 2010.

away from a particular action, e.g., "Do not neglect to order their delicious garlic bread."

- **Questions**, rhetorical or otherwise, can indicate negations often in the context of surprise or bewilderment. For example, a reviewer of a desk phone may write, "Why couldn't they include a decent speaker in this phone?", implying that the phone being reviewed does not have a decent speaker.

- **Supports** and **Repetitions** are used to express agreement and add emphasis or clarity, respectively, and each involve multiple expressions of negation. For the purpose of negation scope detection, each instance of negation in a support or repetition can be isolated and treated as an independent denial or imperative.

Tottie also distinguishes between intersentential and sentential negation. In the case of intersentential negation, the language used in one sentence may explicitly negate a proposition or implication found in another sentence. Rejections and supports are common examples of intersentential negation. Sentential negation, or negations within the scope of a single sentence, are much more frequent; thus sentential denials, imperatives, and questions are the primary focus of the work presented here.

The goal of the present work is to develop a system that is robust to differences in the intended scope of negation introduced by the syntactic and lexical features in each negation category. In particular, as the larger context of this research involves sentiment analysis, it is desirable to construct a negation system that can correctly identify the presence or absence of negation in spans of text that are expressions of sentiment. It so follows that in developing a solution for the specific case of the negation of sentiment, the proposed system is also effective at solving the general case of negation scope identification.

This rest of this paper is organized as follows. §2 presents related work on the topic of automatic detection of the scope of linguistic negations. The annotated corpora used to evaluate the proposed negation scope identification method are presented in §3, including a new data set developed for the purpose of identifying negation scopes in the context of online reviews. §4 describes the proposed negation scope detection system. The novel system is evaluated in §5 in terms of raw results on the annotated negation corpora as well as the performance improvement on sentiment classification achieved by incorporating the negation system in a state-of-the-art sentiment analysis pipeline. Lessons learned and future directions are discussed in §6.

2 Related work

Negation and its scope in the context of sentiment analysis has been studied in the past (Moilanen and Pulman, 2007). In this work we focus on explicit negation mentions, also called functional negation by Choi and Cardie (2008). However, others have studied various forms of negation within the domain of sentiment analysis, including work on content negators, which typically are verbs such as "hampered", "lacked", "denied", etc. (Moilanen and Pulman, 2007; Choi and Cardie, 2008). A recent study by Danescu-Niculescu-Mizil et al. (2009) looked at the problem of finding downward-entailing operators that include a wider range of lexical items, including soft negators such as the adverbs "rarely" and "hardly".

With the absence of a general purpose corpus annotating the precise scope of negation in sentiment corpora, many studies incorporate negation terms through heuristics or soft-constraints in statistical models. In the work of Wilson et al. (2005), a supervised polarity classifier is trained with a set of negation features derived from a list of cue words and a small window around them in the text. Choi and Cardie (2008) combine different kinds of negators with lexical polarity items through various compositional semantic models, both heuristic and machine learned, to improve phrasal sentiment analysis. In that work the scope of negation was either left undefined or determined through surface level syntactic patterns similar to the syntactic patterns from Moilanen and Pulman (2007). A recent study by Nakagawa et al. (2010) developed an semi-supervised model for sub-sentential sentiment analysis that predicts polarity based on the interactions between nodes in dependency graphs, which potentially can induce the scope of negation.

As mentioned earlier, the goal of this work is to define a system that can identify exactly the scope of negation in free text, which requires a robustness to the wide variation of negation expression,

both syntactic and lexical. Thus, this work is complimentary to those mentioned above in that we are measuring not only whether negation detection is useful for sentiment, but to what extent we can determine its exact scope in the text. Towards this end in we describe both an annotated negation span corpus as well as a negation span detector that is trained on the corpus. The span detector is based on conditional random fields (CRFs) (Lafferty, McCallum, and Pereira, 2001), which is a structured prediction learning framework common in sub-sentential natural language processing tasks, including sentiment analysis (Choi and Cardie, 2007; McDonald et al., 2007)

The approach presented here resembles work by Morante and Daelemans (2009), who used IGTree to predict negation cues and a CRF metalearner that combined input from k-nearest neighbor classification, a support vector machine, and another underlying CRF to predict the scope of negations within the BioScope corpus. However, our work represents a simplified approach that replaces machine-learned cue prediction with a lexicon of explicit negation cues, and uses only a single CRF to predict negation scopes, with a more comprehensive model that includes features from a dependency parser.

3 Data sets

One of the only freely available resources for evaluating negation detection performance is the BioScope corpus (Vincze et al., 2008), which consists of annotated clinical radiology reports, biological full papers, and biological abstracts. Annotations in BioScope consist of labeled negation and speculation cues along with the boundary of their associated text scopes. Each cue is associated with exactly one scope, and the cue itself is considered to be part of its own scope. Traditionally, negation detection systems have encountered the most difficulty in parsing the full papers subcorpus, which contains nine papers and a total of 2670 sentences, and so the BioScope full papers were held out as a benchmark for the methods presented here.

The work described in this paper was part of a larger research effort to improve the accuracy of sentiment analysis in online reviews, and it was determined that the intended domain of application would likely contain language patterns that are significantly distinct from patterns common in the text of professional biomedical writings. Correct analysis of reviews generated by web users requires robustness in the face of ungrammatical sentences and misspelling, which are both exceedingly rare in BioScope. Therefore, a novel corpus was developed containing the text of entire reviews, annotated according to spans of negated text.

A sample of 268 product reviews were obtained by randomly sampling reviews from Google Product Search[1] and checking for the presence of negation. The annotated corpus contains 2111 sentences in total, with 679 sentences determined to contain negation. Each review was manually annotated with the scope of negation by a single person, after achieving inter-annotator agreement of 91% with a second person on a smaller subset of 20 reviews containing negation. Inter-annotator agreement was calculated using a strict exact span criteria where both the existence *and* the left/right boundaries of a negation span were required to match. Hereafter the reviews data set will be referred to as the Product Reviews corpus.

The Product Reviews corpus was annotated according to the following instructions:

1. **Negation cues:** Negation cues (e.g., the words "never", "no", or "not" in it's various forms) are not included the negation scope. For example, in the sentence, "It was not X" only "X" is annotated as the negation span.

2. **General Principles:** Annotate the minimal span of a negation covering only the portion of the text being negated semantically. When in doubt, prefer simplicity.

3. **Noun phrases:** Typically entire noun phrases are annotated as within the scope of negation if a noun within the phrase is negated. For example, in the sentence, "This was not a review" the string "a review" is annotated. This is also true for more complex noun phrases, e.g., "This was not a review of a movie that I watched" should be annotated with the span "a review of a movie that I watched".

4. **Adjectives in noun phrases:** Do not annotate an entire noun phrase if an adjective is all that is being negated - consider the negation of each term separately. For instance, "Not

[1] http://www.google.com/products/

top-drawer cinema, but still good...": "top-drawer" is negated, but "cinema" is not, since it is still cinema, just not "top-drawer".

5. **Adverbs/Adjective phrases:**

   (a) Case 1: Adverbial comparatives like "very," "really," "less," "more", etc., annotate the entire adjective phrase, e.g., "It was not very good" should be annotated with the span "very good".

   (b) Case 2: If only the adverb is directly negated, only annotate the adverb itself. E.g., "Not only was it great", or "Not quite as great": in both cases the subject still "is great", so just "only" and "quite" should be annotated, respectively. However, there are cases where the intended scope of adverbial negation is greater, e.g., the adverb phrase "just a small part" in "Tony was on stage for the entire play. It was not just a small part".

   (c) Case 3: "as good as X". Try to identify the intended scope, but typically the entire phrase should be annotated, e.g., "It was not as good as I remember". Note that Case 2 and 3 can be intermixed, e.g., "Not quite as good as I remember", in this case follow 2 and just annotate the adverb "quite", since it was still partly "as good as I remember", just not entirely.

6. **Verb Phrases:** If a verb is directly negated, annotate the entire verb phrase as negated, e.g., "appear to be red" would be marked in "It did not appear to be red".

For the case of verbs (or adverbs), we made no special instructions on how to handle verbs that are content negators. For example, for the sentence "I can't deny it was good", the entire verb phrase "deny it was good" would be marked as the scope of "can't". Ideally annotators would also mark the scope of the verb "deny", effectively canceling the scope of negation entirely over the adjective "good". As mentioned previously, there are a wide variety of verbs and adverbs that play such a role and recent studies have investigated methods for identifying them (Choi and Cardie, 2008; Danescu-Niculescu-Mizil et al., 2009). We leave the identification of the scope of such lexical items and their interaction with explicit negation as future work.

hardly	lack	lacking	lacks
neither	nor	never	no
nobody	none	nothing	nowhere
not	n't	aint	cant
cannot	darent	dont	doesnt
didnt	hadnt	hasnt	havnt
havent	isnt	mightnt	mustnt
neednt	oughtnt	shant	shouldnt
wasnt	wouldnt	without	

Table 1: Lexicon of explicit negation cues.

The Product Reviews corpus is different from BioScope in several ways. First, BioScope ignores direct adverb negation, such that neither the negation cue nor the negation scope in the the phrase, "not only," is annotated in BioScope. Second, BioScope annotations always include entire adjective phrases as negated, where our method distinguishes between the negation of adjectives and adjective targets. Third, BioScope includes negation cues within their negation scopes, whereas our corpus separates the two.

4 System description

As the present work focuses on explicit negations, the choice was made to develop a lexicon of explicit negation cues to serve as primary indicators of the presence of negation. Klima (1964) was the first to identify negation words using a statistics-driven approach, by analyzing word co-occurrence with n-grams that are cues for the presence of negation, such as "either" and "at all". Klima's lexicon served as a starting point for the present work, and was further refined through the inclusion of common misspellings of negation cues and the manual addition of select cues from the "Neg" and "Negate" tags of the General Inquirer (Stone et al., 1966). The final list of cues used for the evaluations in §5 is presented in Table 1. The lexicon serves as a reliable signal to detect the presence of explicit negations, but provides no means of inferring the scope of negation. For scope detection, additional signals derived from surface and dependency level syntactic structure are employed.

The negation scope detection system is built as an individual annotator within a larger annotation pipeline. The negation annotator relies on two dis-

tinct upstream annotators for 1) sentence boundary annotations, derived from a rule-based sentence boundary extractor and 2) token annotations from a dependency parser. The dependency parser is an implementation of the parsing systems described in Nivre and Scholz (2004) and Nivre et al. (2007). Each annotator marks the character offsets for the begin and end positions of individual annotation ranges within documents, and makes the annotations available to downstream processes.

The dependency annotator controls multiple lower-level NLP routines, including tokenization and part of speech (POS) tagging in addition to parsing sentence level dependency structure. The output that is kept for downstream use includes only POS and dependency relations for each token. The tokenization performed at this stage is recycled when learning to identify negation scopes.

The feature space of the learning problem adheres to the dimensions presented in Table 2, and negation scopes are modeled using a first order linear-chain conditional random field (CRF)[2], with a label set of size two indicating whether a token is within or outside of a negation span. The features include the lowercased token string, token POS, token-wise distance from explicit negation cues, POS information from dependency heads, and dependency distance from dependency heads to explicit negation cues. Only unigram features are employed, but each unigram feature vector is expanded to include bigram and trigram representations derived from the current token in conjunction with the prior and subsequent tokens.

The distance measures can be explained as follows. Token-wise distance is simply the number of tokens from one token to another, in the order they appear in a sentence. Dependency distance is more involved, and is calculated as the minimum number of edges that must be traversed in a dependency tree to move from one node (or token) to another. Each edge is considered to be bidirectional. The CRF implementation used in our system employs categorical features, so both integer distances are treated as encodings rather than continuous values. The number 0 implies that a token is, or is part of, an explicit negation cue. The numbers 1-4 encode step-wise distance from a negation cue, and the number 5 is used to jointly encode the concepts of "far away" and "not applicable". The maximum integer distance is 5, which was determined empirically.

[2]Implemented with CRF++: http://crfpp.sourceforge.net/

Feature	Description
Word	The lowercased token string.
POS	The part of speech of a token.
Right Dist.	The linear token-wise distance to the nearest explicit negation cue to the right of a token.
Left Dist.	The linear token-wise distance to the nearest explicit negation cue to the left of a token.
Dep1 POS	The part of speech of the the first order dependency of a token.
Dep1 Dist.	The minimum number of dependency relations that must be traversed to from the first order dependency head of a token to an explicit negation cue.
Dep2 POS	The part of speech of the the second order dependency of a token.
Dep2 Dist.	The minimum number of dependency relations that must be traversed to from the second order dependency head of a token to an explicit negation cue.

Table 2: Token features used in the conditional random field model for negation.

The negation annotator vectorizes the tokens generated in the dependency parser annotator and can be configured to write token vectors to an output stream (training mode) or load a previously learned conditional random field model and apply it by sending the token vectors directly to the CRF decoder (testing mode). The output annotations include document-level negation span ranges as well as sentence-level token ranges that include the CRF output probability vector, as well as the alpha and beta vectors.

5 Results

The negation scope detection system was evaluated against the data sets described in §3. The negation CRF model was trained and tested against the Product Reviews and BioScope biological full papers corpora. Subsequently, the practical effect of robust negation detection was measured in the context of a state-of-the-art sentiment analysis system.

Corpus	Prec.	Recall	F1	PCS
Reviews	81.9	78.2	80.0	39.8
BioScope	80.8	70.8	75.5	53.7

Table 3: Results of negation scope detection.

5.1 Negation Scope Detection

To measure scope detection performance, the automatically generated results were compared against each set of human-annotated negation corpora in a token-wise fashion. That is, precision and recall were calculated as a function of the predicted versus actual class of each text token. Tokens made up purely of punctuation were considered to be arbitrary artifacts of a particular tokenization scheme, and thus were excluded from the results. In keeping with the evaluation presented by Morante and Daelemans (2009), the number of perfectly identified negation scopes is measured separately as the percentage of correct scopes (PCS). The PCS metric is calculated as the number of correct spans divided by the number of true spans, making it a recall measure.

Only binary classification results were considered (whether a token is of class "negated" or "not negated") even though the probabilistic nature of conditional random fields makes it possible to express uncertainty in terms of soft classification scores in the range 0 to 1. Correct predictions of the absence of negation are excluded from the results, so the reported measurements only take into account correct prediction of negation and incorrect predictions of either class.

The negation scope detection results for both the Product Reviews and BioScope corpora are presented in Table 3. The results on the Product Reviews corpus are based on seven-fold cross validation, and the BioScope results are based on five-fold cross validation, since the BioScope data set is smaller. For each fold, the number of sentences with and without negation were balanced in both training and test sets.

The system was designed primarily to support the case of negation scope detection in the open web, and no special considerations were taken to improve performance on the BioScope corpus. In particular, the negation cue lexicon presented in Table 1 was not altered in any way, even though BioScope contains additional cues such as "rather than" and "instead of". This had a noticeable effect on on recall in BioScope, although in several cases the CRF was still able to learn the missing cues indirectly through lexical features.

Condition	Prec.	Recall	F1	PCS
BioScope, trained on Reviews	72.2	42.1	53.5	52.2
Reviews, trained on Bioscope	58.8	68.8	63.4	45.7

Table 4: Results for cross-trained negation models. This shows the results for BioScope with a model trained on the Product Reviews corpus, and the results for Product Reviews with a model trained on the BioScope corpus.

In general, the system performed significantly better on the Product Reviews corpus than on BioScope, although the performance on BioScope full papers is state-of-the-art. This can be accounted for at least partially by the differences in the negation cue lexicons. However, significantly more negation scopes were perfectly identified in BioScope, with a 23% improvement in the PCS metric over the Product Reviews corpus.

The best reported performance to date on the BioScope full papers corpus was presented by Morante and Daelemans (2009), who achieved an F1 score of 70.9 with predicted negation signals, and an F1 score of 84.7 by feeding the manually annotated negation cues to their scope finding system. The system presented here compares favorably to Morante and Daelemans' fully automatic results, achieving an F1 score of 75.5, which is a 15.8% reduction in error, although the results are significantly worse than what was achieved via perfect negation cue information.

5.2 Cross training

The degree to which models trained on each corpus generalized to each other was also measured. For this experiment, each of the two models trained using the methods described in §5.1 was evaluated against its non-corresponding corpus, such that the BioScope-trained corpus was evaluated against all of Product Reviews, and the model derived from Product Reviews was evaluated against all of BioScope.

The cross training results are presented in Table 4. Performance is generally much worse, as expected. Recall drops substantially in BioScope,

which is almost certainly due to the fact that not only are several of the BioScope negation cues missing from the cue lexicon, but the CRF model has not had the opportunity to learn from the lexical features in BioScope. The precision in BioScope remains fairly high, and the percentage of perfectly labeled scopes remains almost the same. For Product Reviews, an opposing trend can be seen: precision drops significantly but recall remains fairly high. This seems to indicate that the scope boundaries in the Product Reviews corpus are generally harder to predict. The percentage of perfectly labeled scopes actually increases for Product Reviews, which could also indicate that scope boundaries are less noisy in BioScope.

5.3 Effect on sentiment classification

In addition to measuring the raw performance of the negation scope detection system, an experiment was conducted to measure the effect of the final negation system within the context of a larger sentiment analysis system.

The negation system was built into a sentiment analysis pipeline consisting of the following stages:

1. Sentence boundary detection.
2. Sentiment detection.
3. Negation scope detection, applying the system described in §4.
4. Sentence sentiment scoring.

The sentiment detection system in stage 2 finds and scores mentions of n-grams found in a large lexicon of sentiment terms and phrases. The sentiment lexicon is based on recent work using label propagation over a very large distributional similarity graph derived from the web (Velikovich et al., 2010), and applies positive or negative scores to terms such as "good", "bad", or "just what the doctor ordered". The sentence scoring system in stage 4 then determines whether any scored sentiment terms fall within the scope of a negation, and flips the sign of the sentiment score for all negated sentiment terms. The scoring system then sums all sentiment scores within each sentence and computes overall sentence sentiment scores.

A sample of English-language online reviews was collected, containing a total of 1135 sentences. Human raters were presented with consecutive sentences and asked to classify each sentence as expressing one of the following types of sentiment: 1) positive, 2) negative, 3) neutral, or 4) mixed positive and negative. Each sentence was reviewed independently by five separate raters, and final sentence classification was determined by consensus. Of the original 1135 sentences 216, or 19%, were found to contain negations.

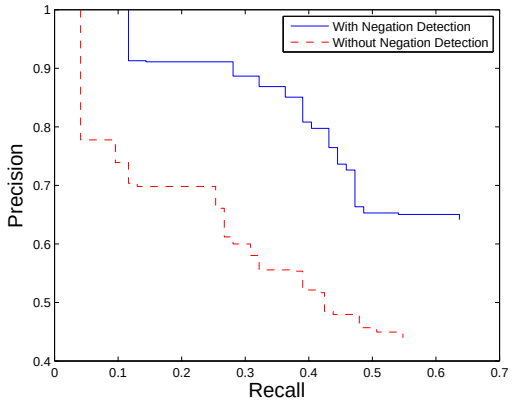

Figure 1: Precision-recall curve showing the effect of negation detection on positive sentiment prediction.

The effect of the negation system on sentiment classification was evaluated on the smaller subset of 216 sentences in order to more precisely measure the impact of negation detection. The smaller negation subset contained 73 sentences classified as positive, 114 classified as negative, 12 classified as neutral, and 17 classified as mixed. The number of sentences classified as neutral or mixed was too small for a useful performance measurement, so only sentences classified as positive or negative sentences were considered.

Figures 1 and 2 show the precision-recall curves for sentences predicted by the sentiment analysis system to be positive and negative, respectively. The curves indicate relatively low performance, which is consistent with the fact that sentiment polarity detection is notoriously difficult on sentences with negations. The solid lines show performance with the negation scope detection system in place, and the dashed lines show performance with no negation detection at all. From the figures, a significant improvement is immediately apparent at all recall levels. It can also be inferred from the figures that the sentiment analysis system is significantly biased towards positive predictions: even though there were significantly more sentences classified by human raters as neg-

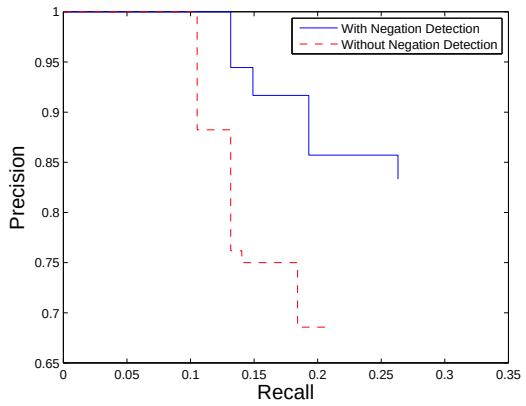

Figure 2: Precision-recall curve showing the effect of negation detection on negative sentiment prediction.

Metric	w/o Neg.	w/ Neg.	% Improv.
Positive Sentiment			
Prec.	44.0	64.1	35.9
Recall	54.8	63.7	20.0
F1	48.8	63.9	29.5
Negative Sentiment			
Prec.	68.6	83.3	46.8
Recall	21.1	26.3	6.6
F1	32.3	40.0	11.4

Table 5: Sentiment classification results, showing the percentage improvement obtained from including negation scope detection (w/ Neg.) over results obtained without including negation scope detection (w/o Neg.).

ative, the number of data points for positive predictions far exceeds the number of negative predictions, with or without negation detection.

The overall results are presented in Table 5, separated by positive and negative class predictions. As expected, performance is improved dramatically by introducing negation scope detection. The precision of positive sentiment predictions sees the largest improvement, largely due to the inherent bias in the sentiment scoring algorithm. F1 scores for positive and negative sentiment predictions improve by 29.5% and 11.4%, respectively.

6 Conclusions

This paper presents a system for identifying the scope of negation using shallow parsing, by means of a conditional random field model informed by a dependency parser. Results were presented on the standard BioScope corpus that compare favorably to the best results reported to date, using a software stack that is significantly simpler than the best-performing approach.

A new data set was presented that targets the domain of online product reviews. The product review corpus represents a departure from the standard BioScope corpus in two distinct dimensions: the reviews corpus contains diverse common and vernacular language patterns rather than professional prose, and also presents a divergent method for annotating negations in text. Cross-training by learning a model on one corpus and testing on another suggests that scope boundary detection in the product reviews corpus may be a more difficult learning problem, although the method used to annotate the reviews corpus may result in a more consistent representation of the problem.

Finally, the negation system was built into a state-of-the-art sentiment analysis system in order to measure the practical impact of accurate negation scope detection, with dramatic results. The negation system improved the precision of positive sentiment polarity detection by 35.9% and negative sentiment polarity detection by 46.8%. Error reduction on the recall measure was less dramatic, but still significant, showing improved recall for positive polarity of 20.0% and improved recall for negative polarity of 6.6%.

Future research will include treatment of implicit negation cues, ideally by learning to predict the presence of implicit negation using a probabilistic model that generates meaningful confidence scores. A related topic to be addressed is the automatic detection of sarcasm, which is an important problem for proper sentiment analysis, particularly in open web domains where language is vernacular. Additionally, we would like to tackle the problem of inter-sentential negations, which could involve a natural extension of negation scope detection through co-reference resolution, such that negated pronouns trigger negations in text surrounding their pronoun antecedents.

Acknowledgments

The authors would like to thank Andrew Hogue and Kerry Hannan for useful discussions regarding this work.

References

Yejin Choi and Claire Cardie. 2007. Structured Local Training and Biased Potential Functions for Conditional Random Fields with Application to Coreference Resolution. *Proceedings of The 9th Conference of the North American Chapter of the Association for Computational Linguistics,* ACL, Rochester, NY.

Yejin Choi and Claire Cardie. 2008. Learning with Compositional Semantics as Structural Inference for Subsentential Sentiment Analysis. *Proceedings of the Conference on Empirical Methods on Natural Language Processing.* ACL, Honolulu, HI.

Cristian Danescu-Niculescu-Mizil, Lillian Lee, and Richard Ducott. 2008. Without a 'doubt'? Unsupervised discovery of downward-entailing operators. *Proceedings of The 10th Annual Conference of the North American Chapter of the Association for Computational Linguistics.* ACL, Boulder, CO.

Talmy Givón. 1993. *English Grammer: A Function-Based Introduction.* Benjamins, Amsterdam, NL.

Edward S. Klima. 1964. Negation in English. *Readings in the Philosophy of Language.* Ed. J. A. Fodor and J. J. Katz. Prentice Hall, Englewood Cliffs, NJ: 246-323.

John Lafferty, Andrew McCallum, and Fernando Pereira. 2001. Conditional random elds: Probabilistic models for segmenting and labeling sequence data. *Proceedings of the International Conference on Machine Learning*. Morgan Kaufmann, Williamstown, MA.

Ryan McDonald, Kerry Hannan, Tyler Neylon, Mike Wells, and Jeff Reynar. 2007. Structured Models for Fine-to-Coarse Sentiment Analysis. *Proceedings of the Annual Meeting of the Association for Computational Linguistics.* Prague, Czech Republic.

Karo Moilanen and Stephen Pulman 2007. Sentiment Composition. *Proceedings of the Recent Advances in Natural Language Processing International Conference* Borovets, Bulgaria

Roser Morante and Walter Daelemans. 2009. A metalearning approach to processing the scope of negation. *Proceedings of the Thirteenth Conference on Computational Natural Language Learning (CoNLL).* ACM, Boulder, CO.

Tetsuji Nakagawa, Kentaro Inui, and Sadao Kurohashi. 2010. Dependency Tree-based Sentiment Classification using CRFs with Hidden Variables. *Proceedings of The 11th Annual Conference of the North American Chapter of the Association for Computational Linguistics* ACL, Los Angeles, CA.

Joakim Nivre and Mario Scholz. 2004. Deterministic Dependency Parsing of English Text. *Proceedings of the 20th International Conference on Computational Linguistics.* ACM, Geneva, Switzerland.

Joakim Nivre, Johan Hall, Jens Nilsson, Atanas Chanev, Gulsen Eryigit Sandra Kubler, Svetoslav Marinov and Erwin Marsi. 2007. MaltParser: A language-independent system for data-driven dependency parsing *Natural Language Engineering* 13(02):95–135

Philip J. Stone, Dexter C. Dunphy, Marshall S. Smith, Daniel M. Ogilvie. 1966. *The General Inquirer: A Computer Approach to Content Analysis.* MIT Press, Cambridge, MA.

Gunnel Tottie. 1991. *Negation in English Speech and Writing: A Study in Variation* Academic, San Diego, CA.

Leonid Velikovich, Sasha Blair-Goldensohn, Kerry Hannan, and Ryan McDonald. 2010. The viability of web-derived polarity lexicons. *Proceedings of The 11th Annual Conference of the North American Chapter of the Association for Computational Linguistics.* ACL, Los Angeles, CA.

Veronika Vincze, György Szarvas, Richárd Farkas, György Móra, and János Csirik. 2008. The BioScope corpus: biomedical texts annotated for uncertainty, negation and their scopes. *BMC Bioinformatics*, 9(Suppl 11):S9.

Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. 2005. Recognizing contextual polarity in phrase-level sentiment analysis. *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing* Vancouver, Canada.

A Survey on the Role of Negation in Sentiment Analysis

Michael Wiegand
Saarland University
Saarbrücken, Germany
michael.wiegand@lsv.uni-saarland.de

Alexandra Balahur
University of Alicante
Alicante, Spain
abalahur@dlsi.ua.es

Benjamin Roth and **Dietrich Klakow**
Saarland University
Saarbrücken, Germany
benjamin.roth@lsv.uni-saarland.de
dietrich.klakow@lsv.uni-saarland.de

Andrés Montoyo
University of Alicante
Alicante, Spain
montoyo@dlsi.ua.es

Abstract

This paper presents a survey on the role of *negation* in sentiment analysis. Negation is a very common linguistic construction that affects polarity and, therefore, needs to be taken into consideration in sentiment analysis.

We will present various computational approaches modeling negation in sentiment analysis. We will, in particular, focus on aspects, such as level of representation used for sentiment analysis, negation word detection and scope of negation. We will also discuss limits and challenges of negation modeling on that task.

1 Introduction

Sentiment analysis is the task dealing with the automatic detection and classification of opinions expressed in text written in natural language.

Subjectivity is defined as the linguistic expression of somebody's opinions, sentiments, emotions, evaluations, beliefs and speculations (Wiebe, 1994). Subjectivity is opposed to objectivity, which is the expression of facts. It is important to make the distinction between subjectivity detection and sentiment analysis, as they are two separate tasks in natural language processing. Sentiment analysis can be dependently or independently done from subjectivity detection, although Pang and Lee (2004) state that subjectivity detection performed prior to the sentiment analysis leads to better results in the latter.

Although research in this area has started only recently, the substantial growth in subjective information on the world wide web in the past years has made sentiment analysis a task on which constantly growing efforts have been concentrated. The body of research published on sentiment analysis has shown that the task is difficult, not only due to the syntactic and semantic variability of language, but also because it involves the extraction of indirect or implicit assessments of objects, by means of emotions or attitudes. Being a part of subjective language, the expression of opinions involves the use of nuances and intricate surface realizations. That is why the automatic study of opinions requires fine-grained linguistic analysis techniques and substantial efforts to extract features for machine learning or rule-based systems, in which subtle phenomena as *negation* can be appropriately incorporated.

Sentiment analysis is considered as a subsequent task to subjectivity detection, which should ideally be performed to extract content that is not factual in nature. Subsequently, sentiment analysis aims at classifying the sentiment of the opinions into polarity types (the common types are positive and negative). This text classification task is also referred to as *polarity classification*.

This paper presents a survey on the role of *negation* in sentiment analysis. Negation is a very common linguistic construction that affects polarity and, therefore, needs to be taken into consideration in sentiment analysis. Before we describe the computational approaches that have been devised to account for this phenomenon in sentiment analysis, we will motivate the problem.

2 Motivation

Since subjectivity and sentiment are related to expressions of personal attitudes, the way in which this is realized at the surface level influences the manner in which an opinion is extracted and its polarity is computed. As we have seen, sentiment analysis goes a step beyond subjectivity detection,

Proceedings of the Workshop on Negation and Speculation in Natural Language Processing, pages 60–68,
Uppsala, July 2010.

including polarity classification. So, in this task, correctly determining the valence of a text span (whether it conveys a positive or negative opinion) is equivalent to the success or failure of the automatic processing.
It is easy to see that Sentence 1 expresses a positive opinion.

1. I *like*$^{+}$ this new Nokia model.

The polarity is conveyed by *like* which is a *polar expression*. Polar expressions, such as *like* or *horrible*, are words containing a prior polarity. The negation of Sentence 1, i.e. Sentence 2, using the negation word *not*, expresses a negative opinion.

2. I do [*not* like^{+}]$^{-}$ this new Nokia model.

In this example, it is straightforward to notice the impact of negation on the polarity of the opinion expressed. However, it is not always that easy to spot positive and negative opinions in text. A negation word can also be used in other expressions without constituting a negation of the proposition expressed as exemplified in Sentence 3.

3. *Not only* is this phone expensive *but* it is *also* heavy and difficult to use.

In this context, *not* does not invert the polarity of the opinion expressed which remains negative.
Moreover, the presence of an actual negation word in a sentence does not mean that all its polar opinions are inverted. In Sentence 4, for example, the negation does not modify the second polar expression *intriguing* since the negation and *intriguing* are in separate clauses.

4. [I do [*not* like^{+}]$^{-}$ the design of new Nokia model] but [it contains some intriguing^{+} new functions].

Therefore, when treating negation, one must be able to correctly determine the scope that it has (i.e. determine what part of the meaning expressed is modified by the presence of the negation).
Finally, the surface realization of a negation is highly variable, depending on various factors, such as the impact the author wants to make on the general text meaning, the context, the textual genre etc. Most of the times, its expression is far from being simple (as in the first two examples), and does not only contain obvious negation words, such as *not*, *neither* or *nor*. Research in the field has shown that there are many other words that invert the polarity of an opinion expressed, such as *diminishers*/*valence shifters* (Sentence 5), *connectives* (Sentence 6), or even *modals* (Sentence 7).

5. I find the functionality of the new phone *less* practical.

6. Perhaps it is a great phone, *but* I fail to see why.

7. In theory, the phone *should* have worked even under water.

As can be seen from these examples, modeling negation is a difficult yet important aspect of sentiment analysis.

3 The Survey

In this survey, we focus on work that has presented novel aspects for negation modeling in sentiment analysis and we describe them chronologically.

3.1 Negation and Bag of Words in Supervised Machine Learning

Several research efforts in polarity classification employ supervised machine-learning algorithms, like Support Vector Machines, Naïve Bayes Classifiers or Maximum Entropy Classifiers. For these algorithms, already a low-level representation using bag of words is fairly effective (Pang et al., 2002). Using a bag-of-words representation, the supervised classifier has to figure out by itself which words in the dataset, or more precisely feature set, are polar and which are not. One either considers all words occurring in a dataset or, as in the case of Pang et al. (2002), one carries out a simple feature selection, such as removing infrequent words. Thus, the standard bag-of-words representation does not contain any explicit knowledge of polar expressions. As a consequence of this simple level of representation, the reversal of the polarity type of polar expressions as it is caused by a negation cannot be explicitly modeled. The usual way to incorporate negation modeling into this representation is to add artificial words: i.e. if a word *x* is preceded by a negation word, then rather than considering this as an occurrence of the feature *x*, a new feature *NOT_x* is created. The scope of negation cannot be properly modeled with this representation either. Pang et al. (2002), for example, consider every word until the next punctuation mark. Sentence 2 would, therefore, result in the following representation:

8. I do not NOT_like NOT_this NOT_new NOT_Nokia NOT_model.

The advantage of this feature design is that a plain occurrence and a negated occurrence of a word are

reflected by two separate features. The disadvantage, however, is that these two contexts treat the same word as two completely different entities.
Since the words to be considered are unrestricted, any word – no matter whether it is an actual polar expression or not – is subjected to this negation modification. This is not only linguistically inaccurate but also increases the feature space with more sparse features (since the majority of words will only be negated once or twice in a corpus).
Considering these shortcomings, it comes to no surprise that the impact of negation modeling on this level of representation is limited. Pang et al. (2002) report only a *negligible* improvement by adding the artificial features compared to plain bag of words in which negation is not considered.
Despite the lack of linguistic plausibility, supervised polarity classifiers using bag of words (in particular, if training and testing are done on the same domain) offer fairly good performance. This is, in particular, the case on coarse-grained classification, such as on document level. The success of these methods can be explained by the fact that larger texts contain redundant information, e.g. it does not matter whether a classifier cannot model a negation if the text to be classified contains twenty polar opinions and only one or two contain a negation. Another advantage of these machine learning approaches on coarse-grained classification is their usage of higher order *n-grams*. Imagine a labeled training set of documents contains frequent bigrams, such as *not appealing* or *less entertaining*. Then a feature set using higher order n-grams implicitly contains negation modeling. This also partially explains the effectiveness of bigrams and trigrams for this task as stated in (Ng et al., 2006).
The dataset used for the experiments in (Pang et al., 2002; Ng et al., 2006) has been established as a popular benchmark dataset for sentiment analysis and is publicly available[1].

3.2 Incorporating Negation in Models that Include Knowledge of Polar Expressions - Early Works

The previous subsection suggested that appropriate negation modeling for sentiment analysis requires the awareness of polar expressions. One way of obtaining such expressions is by using a polarity lexicon which contains a list of polar expressions and for each expression the corresponding polarity type. A simple rule-based polarity classifier derived from this knowledge typically counts the number of positive and negative polar expressions in a text and assigns it the polarity type with the majority of polar expressions. The counts of polar expressions can also be used as features in a supervised classifier. Negation is typically incorporated in those features, e.g. by considering negated polar expressions as unnegated polar expressions with the opposite polarity type.

3.2.1 Contextual Valence Shifters

The first computational model that accounts for negation in a model that includes knowledge of polar expressions is (Polanyi and Zaenen, 2004). The different types of negations are modeled via *contextual valence shifting*. The model assigns scores to polar expressions, i.e. positive scores to positive polar expressions and negative scores to negative polar expressions, respectively. If a polar expression is negated, its polarity score is simply inverted (see Example 1).

$$\text{clever } (+2) \rightarrow \text{not clever } (-2) \tag{1}$$

In a similar fashion, diminishers are taken into consideration. The difference is, however, that the score is only reduced rather than shifted to the other polarity type (see Example 2).

$$\text{efficient } (+2) \rightarrow \text{rather efficient } (+1) \tag{2}$$

Beyond that the model also accounts for modals, presuppositional items and even discourse-based valence shifting. Unfortunately, this model is not implemented and, therefore, one can only speculate about its real effectiveness.

Kennedy and Inkpen (2005) evaluate a negation model which is fairly identical to the one proposed by Polanyi and Zaenen (2004) (as far as simple negation words and diminishers are concerned) in document-level polarity classification. A simple scope for negation is chosen. A polar expression is thought to be negated if the negation word immediately precedes it. In an extension of this work (Kennedy and Inkpen, 2006) a parser is considered for scope computation. Unfortunately, no precise description of how the parse is used for scope modeling is given in that work. Neither is there a comparison of these two scope models measuring their respective impacts.

[1] http://www.cs.cornell.edu/people/pabo/movie-review-data

Final results show that modeling negation is important and relevant, even in the case of such simple methods. The consideration of negation words is more important than that of diminishers.

3.2.2 Features for Negation Modeling

Wilson et al. (2005) carry out more advanced negation modeling on expression-level polarity classification. The work uses supervised machine learning where negation modeling is mostly encoded as features using polar expressions. The features for negation modeling are organized in three groups:

- negation features
- shifter features
- polarity modification features

Negation features directly relate to negation expressions negating a polar expression. One feature checks whether a negation expression occurs in a fixed window of four words preceding the polar expression. The other feature accounts for a polar predicate having a negated subject. This frequent long-range relationship is illustrated in Sentence 9.

9. [No politically prudent Israeli]$_{subject}$ could support$_{polar\ pred}$ either of them.

All negation expressions are additionally disambiguated as some negation words do not function as a negation word in certain contexts, e.g. *not to mention* or *not just*.

Shifter features are binary features checking the presence of different types of *polarity shifters*. Polarity shifters, such as *little*, are weaker than ordinary negation expressions. They can be grouped into three categories, general polarity shifters, positive polarity shifters, and negative polarity shifters. General polarity shifters reverse polarity like negations. The latter two types only reverse a particular polarity type, e.g. the positive shifter *abate* only modifies negative polar expressions as in *abate the damage*. Thus, the presence of a positive shifter may indicate positive polarity. The set of words that are denoted by these three features can be approximately equated with diminishers.

Finally, *polarity modification features* describe polar expressions of a particular type modifying or being modified by other polar expressions. Though these features do not explicitly contain negations, language constructions which are similar to negation may be captured. In the phrase [*disappointed*$^{-}$ *hope*$^{+}$]$^{-}$, for instance, a negative polar expression modifies a positive polar expression which results in an overall negative phrase.

Adding these three feature groups to a feature set comprising bag of words and features counting polar expressions results in a significant improvement. In (Wilson et al., 2009), the experiments of Wilson et al. (2005) are extended by a detailed analysis on the individual effectiveness of the three feature groups mentioned above. The results averaged over four different supervised learning algorithms suggest that the actual negation features are most effective whereas the binary polarity shifters have the smallest impact. This is consistent with Kennedy and Inkpen (2005) given the similarity of polarity shifters and diminishers.

Considering the amount of improvement that is achieved by negation modeling, the improvement seems to be larger in (Wilson et al., 2005). There might be two explanations for this. Firstly, the negation modeling in (Wilson et al., 2005) is considerably more complex and, secondly, Wilson et al. (2005) evaluate on a more fine-grained level (i.e. expression level) than Kennedy and Inkpen (2005) (they evaluate on document level). As already pointed out in §3.1, document-level polarity classification contains more redundant information than sentence-level or expression-level polarity classification, therefore complex negation modeling on these levels might be more effective since the correct contextual interpretation of an individual polar expression is far more important[2].

The fine-grained opinion corpus used in (Wilson et al., 2005; Wilson et al., 2009) and all the resources necessary to replicate the features used in these experiments are also publicly available[3].

3.3 Other Approaches

The approaches presented in the previous section (Polanyi and Zaenen, 2004; Kennedy and Inkpen, 2005; Wilson et al., 2005) can be considered as the works pioneering negation modeling in sentiment analysis. We now present some more recent work on that topic. All these approaches, however, are heavily related to these early works.

[2] This should also explain why most subsequent works (see §3.3) have been evaluated on fine-grained levels.

[3] The corpus is available under: `http://www.cs.pitt.edu/mpqa/databaserelease` and the resources for the features are part of OpinionFinder: `http://www.cs.pitt.edu/mpqa/opinionfinderrelease`

3.3.1 Semantic Composition

In (Moilanen and Pulman, 2007), a method to compute the polarity of headlines and complex noun phrases using compositional semantics is presented. The paper argues that the principles of this linguistic modeling paradigm can be successfully applied to determine the subsentential polarity of the sentiment expressed, demonstrating it through its application to contexts involving sentiment propagation, polarity reversal (e.g. through the use of negation following Polanyi and Zaenen (2004) and Kennedy and Inkpen (2005)) or polarity conflict resolution. The goal is achieved through the use of syntactic representations of sentences, on which rules for composition are defined, accounting for negation (incrementally applied to constituents depending on the scope) using negation words, shifters and negative polar expressions. The latter are subdivided into different categories, such that special words are defined, whose negative intensity is strong enough that they have the power to change the polarity of the entire text spans or constituents they are part of.

A similar approach is presented by Shaikh et al. (2007). The main difference to Moilanen and Pulman (2007) lies in the representation format on which the compositional model is applied. While Moilanen and Pulman (2007) use syntactic phrase structure trees, Shaikh et al. (2007) consider a more abstract level of representation being verb frames. The advantage of a more abstract level of representation is that it more accurately represents the meaning of the text it describes. Apart from that, Shaikh et al. (2007) design a model for sentence-level classification rather than for headlines or complex noun phrases.

The approach by Moilanen and Pulman (2007) is not compared against another established classification method whereas the approach by Shaikh et al. (2007) is evaluated against a non-compositional rule-based system which it outperforms.

3.3.2 Shallow Semantic Composition

Choi and Cardie (2008) present a more lightweight approach using compositional semantics towards classifying the polarity of expressions. Their working assumption is that the polarity of a phrase can be computed in two steps:

- the assessment of polarity of the constituents
- the subsequent application of a set of previously-defined inference rules

An example rule, such as:

$$Polarity([\text{NP1}]^{-}\ [\text{IN}]\ [\text{NP2}]^{-}) = + \quad (3)$$

may be applied to expressions, such as $[\textit{lack}]^{-}_{\text{NP1}}\ [\textit{of}]_{\text{IN}}\ \ [\textit{crime}]^{-}_{\text{NP2}}$ *in rural areas*. The advantage of these rules is that they restrict the scope of negation to specific constituents rather than using the scope of the entire target expression.

Such inference rules are very reminiscent of *polarity modification features* (Wilson et al., 2005), as a negative polar expression is modified by positive polar expression. The rules presented by Choi and Cardie (2008) are, however, much more specific, as they define syntactic contexts of the polar expressions. Moreover, from each context a direct polarity for the entire expression can be derived. In (Wilson et al., 2005), this decision is left to the classifier. The rules are also similar to the syntactic rules from Moilanen and Pulman (2007). However, they involve less linguistic processing and are easier to comprehend[4]. The effectiveness of these rules are both evaluated in rule-based methods and a machine learning based method where they are anchored as constraints in the objective function. The results of their evaluation show that the compositional methods outperform methods using simpler scopes for negation, such as considering the scope of the entire target expression. The learning method incorporating the rules also slightly outperforms the (plain) rule-based method.

3.3.3 Scope Modeling

In sentiment analysis, the most prominent work examining the impact of different scope models for negation is (Jia et al., 2009). The scope detection method that is proposed considers:

- static delimiters
- dynamic delimiters
- heuristic rules focused on polar expressions

Static delimiters are unambiguous words, such as *because* or *unless* marking the beginning of another clause. *Dynamic delimiters* are, however,

[4]It is probably due to the latter, that these rules have been successfully re-used in subsequent works, most prominently Klenner et al. (2009).

ambiguous, e.g. *like* and *for*, and require disambiguation rules, using contextual information such as their pertaining part-of-speech tag. These delimiters suitably account for various complex sentence types so that only the clause containing the negation is considered.

The *heuristic rules* focus on cases in which polar expressions in specific syntactic configurations are directly preceded by negation words which results in the polar expression becoming a delimiter itself. Unlike Choi and Cardie (2008), these rules require a proper parse and reflect grammatical relationships between different constituents.

The complexity of the scope model proposed by Jia et al. (2009) is similar to the ones of the compositional models (Moilanen and Pulman, 2007; Shaikh et al., 2007; Choi and Cardie, 2008) where scope modeling is exclusively incorporated in the compositional rules.

Apart from scope modeling, Jia et al. (2009) also employ a complex negation term disambiguation considering not only phrases in which potential negation expressions do not have an actual negating function (as already used in (Wilson et al., 2005)), but also *negative rhetorical questions* and *restricted comparative sentences*.

On sentence-level polarity classification, their scope model is compared with

- a simple negation scope using a fixed window size (similar to the negation feature in (Wilson et al., 2005))
- the text span until the first occurrence of a polar expression following the negation word
- the entire sentence

The proposed method consistently outperforms the simpler methods proving that the incorporation of linguistic insights into negation modeling is meaningful. Even on polarity document retrieval, i.e. a more coarse-grained classification task where contextual disambiguation usually results in a less significant improvement, the proposed method also outperforms the other scopes examined.

There have only been few research efforts in sentiment analysis examining the impact of scope modeling for negation in contrast to other research areas, such as the biomedical domain (Huang and Lowe, 2007; Morante et al., 2008; Morante and Daelemans, 2009). This is presumably due to the fact that only for the biomedical domain, publicly available corpora containing annotation for the scope of negation exist (Szarvas et al., 2008). The usability of those corpora for sentiment analysis has not been tested.

3.4 Negation within Words

So far, negation has only be considered as a phenomenon that affects entire words or phrases. The word expressing a negation and the words or phrases being negated are disjoint. There are, however, cases in which both negation and the negated content which can also be opinionated are part of the same word. In case, these words are lexicalized, such as *flaw-less*, and are consequently to be found a polarity lexicon, this phenomenon does not need to be accounted for in sentiment analysis. However, since this process is (at least theoretically) productive, fairly uncommon words, such as *not-so-nice*, *anti-war* or *offensive-less* which are not necessarily contained in lexical resources, may emerge as a result of this process. Therefore, a polarity classifier should also be able to decompose words and carry out negation modeling within words.

There are only few works addressing this particular aspect (Moilanen and Pulman, 2008; Ku et al., 2009) so it is not clear how much impact this type of negation has on an overall polarity classification and what complexity of morphological analysis is really necessary. We argue, however, that in synthetic languages where negation may regularly be realized as an affix rather than an individual word, such an analysis is much more important.

3.5 Negation in Various Languages

Current research in sentiment analysis mainly focuses on English texts. Since there are significant structural differences among the different languages, some particular methods may only capture the idiosyncratic properties of the English language. This may also affect negation modeling.

The previous section already stated that the need for morphological analyses may differ across the different languages.

Moreover, the complexity of scope modeling may also be language dependent. In English, for example, modeling the scope of a negation as a fixed window size of words following the occurrence of a negation expression already yields a reasonable performance (Kennedy and Inkpen, 2005). However, in other languages, for example German, more complex processing is required as the negated expression may either precede (Sen-

tence 10) or follow (Sentence 11) the negation expression. Syntactic properties of the negated noun phrase (i.e. the fact whether the negated polar expression is a verb or an adjective) determine the particular negation construction.

10. *Peter mag den Kuchen nicht.*
    Peter likes the cake not.
    'Peter does not like the cake.'

11. *Der Kuchen ist nicht köstlich.*
    The cake is not delicious.
    'The cake is not delicious.'

These items show that, clearly, some more extensive cross-lingual examination is required in order to be able to make statements of the general applicability of specific negation models.

3.6 *Bad* and *Not Good* are Not the Same

The standard approach of negation modeling suggests to consider a negated polar expression, such as *not bad*, as an unnegated polar expression with the opposite polarity, such as *good*. Liu and Seneff (2009) claim, however, that this is an oversimplification of language. *Not bad* and *good* may have the same polarity but they differ in their respective polar strength, i.e. *not bad* is less positive than *good*. That is why, Liu and Seneff (2009) suggest a compositional model in which for individual adjectives and adverbs (the latter include negations) a prior rating score encoding their intensity and polarity is estimated from pros and cons of on-line reviews. Moreover, compositional rules for polar phrases, such as *adverb-adjective* or *negation-adverb-adjective* are defined exclusively using the scores of the individual words. Thus, adverbs function like universal quantifiers scaling either up or down the polar strength of the specific polar adjectives they modify. The model independently learns what negations are, i.e. a subset of adverbs having stronger negative scores than other adverbs. In short, the proposed model provides a unifying account for intensifiers (e.g. *very*), diminishers, polarity shifters and negation words. Its advantage is that polarity is treated compositionally and is interpreted as a continuum rather than a binary classification. This approach reflects its meaning in a more suitable manner.

3.7 Using Negations in Lexicon Induction

Many classification approaches illustrated above depend on the knowledge of which natural language expressions are polar. The process of acquiring such lexical resources is called lexicon induction. The observation that negations co-occur with polar expressions has been used for inducing polarity lexicons on Chinese in an unsupervised manner (Zagibalov and Carroll, 2008). One advantage of negation is that though the induction starts with just positive polar seeds, the method also accomplishes to extract negative polar expressions since negated mentions of the positive polar seeds co-occur with negative polar expressions. Moreover, and more importantly, the distribution of the co-occurrence between polar expressions and negations can be exploited for the selection of those seed lexical items. The model presented by Zagibalov and Carroll (2008) relies on the observation that a polar expression can be negated but it occurs more frequently without the negation. The distributional behaviour of an expression, i.e. significantly often co-occurring with a negation word but significantly more often occurring without a negation word makes up a property of a polar expression. The data used for these experiments are publicly available[5].

3.8 Irony – The Big Challenge

Irony is a rhetorical process of intentionally using words or expressions for uttering meaning that is different from the one they have when used literally (Carvalho et al., 2009). Thus, we consider that the use of irony can reflect an implicit negation of what is conveyed through the literal use of the words. Moreover, due to its nature irony is mostly used to express a polar opinion.

Carvalho et al. (2009) confirm the relevance of (verbal) irony for sentiment analysis by an error analysis of their present classifier stating that a large proportion of misclassifications derive from their system's inability to account for irony.

They present predictive features for detecting irony in positive sentences (which are actually meant to have a negative meaning). Their findings are that the use of emoticons or expressions of gestures and the use of quotation marks within a context in which no reported speech is included are a good signal of irony in written text. Although the use of these clues in the defined patterns helps to detect some situations in which irony is present, they do not fully represent the phenomenon.

[5] `http://www.informatics.sussex.ac.uk/users/tz21/coling08.zip`

A data-driven approach for irony detection on product-reviews is presented in (Tsur et al., 2010). In the first stage, a considerably large list of simple surface patterns of ironic expressions are induced from a small set of labeled seed sentences. A pattern is a generalized word sequence in which content words are replaced by a generic *CW* symbol. In the second stage, the seed sentences are used to collect more examples from the web, relying on the assumption that sentences next to ironic ones are also ironic. In addition to these patterns, some punctuation-based features are derived from the labeled sentences. The acquired patterns are used as features along the punctuation-based features within a k nearest neighbour classifier. On an in-domain test set the classifier achieves a reasonable performance. Unfortunately, these experiments only elicit few additional insights into the general nature of irony. As there is no cross-domain evaluation of the system, it is unclear in how far this approach generalizes to other domains.

4 Limits of Negation Modeling in Sentiment Analysis

So far, this paper has not only outlined the importance of negation modeling in sentiment analysis but it has also shown different ways to account for this linguistic phenomenon. In this section, we present the limits of negation modeling in sentiment analysis.

Earlier in this paper, we stated that negation modeling depends on the knowledge of polar expressions. However, the recognition of genuine polar expressions is still fairly brittle. Many polar expressions, such as *disease* are ambiguous, i.e. they have a polar meaning in one context (Sentence 12) but do not have one in another (Sentence 13).

12. He is a *disease* to every team he has gone to.

13. Early symptoms of the *disease* are headaches, fevers, cold chills and body pain.

In a pilot study (Akkaya et al., 2009), it has already been shown that applying *subjectivity word sense disambiguation* in addition to the feature-based negation modeling approach of Wilson et al. (2005) results in an improvement of performance in polarity classification.

Another problem is that some polar opinions are not lexicalized. Sentence 14 is a negative *pragmatic opinion* (Somasundaran and Wiebe, 2009) which can only be detected with the help of external world knowledge.

14. The next time I hear this song on the radio, I'll throw my radio out of the window.

Moreover, the effectiveness of specific negation models can only be proven with the help of corpora containing those constructions or the type of language behaviour that is reflected in the models to be evaluated. This presumably explains why rare constructions, such as negations using connectives (Sentence 6 in §2), modals (Sentence 7 in §2) or other phenomena presented in the conceptual model of Polanyi and Zaenen (2004), have not yet been dealt with.

5 Conclusion

In this paper, we have presented a survey on the role of negation in sentiment analysis. The plethora of work presented on the topic proves that this common linguistic construction is highly relevant for sentiment analysis.

An effective negation model for sentiment analysis usually requires the knowledge of polar expressions. Negation is not only conveyed by common negation words but also other lexical units, such as diminishers. Negation expressions are ambiguous, i.e. in some contexts do not function as a negation and, therefore, need to be disambiguated. A negation does not negate every word in a sentence, therefore, using syntactic knowledge to model the scope of negation expressions is useful.

Despite the existence of several approaches to negation modeling for sentiment analysis, in order to make general statements about the effectiveness of specific methods systematic comparative analyses examining the impact of different negation models (varying in complexity) with regard to classification type, text granularity, target domain, language etc. still need to be carried out.

Finally, negation modeling is only one aspect that needs to be taken into consideration in sentiment analysis. In order to fully master this task, other aspects, such as a more reliable identification of genuine polar expressions in specific contexts, are at least as important as negation modeling.

Acknowledgements

Michael Wiegand was funded by the BMBF project NL-Search under contract number 01IS08020B. Alexandra Balahur was funded by Ministerio de Ciencia e Innovación - Spanish Government (grant no. TIN2009-13391-C04-01), and Conselleria d'Educación-Generalitat Valenciana (grant no. PROMETEO/2009/119 and ACOMP/2010/286).

References

C. Akkaya, J. Wiebe, and R. Mihalcea. 2009. Subjectivity Word Sense Disambiguation. In *Proceedings of EMNLP*.

P. Carvalho, L. Sarmento, M. J. Silva, and E. de Oliveira. 2009. Clues for Detecting Irony in User-Generated Contents: Oh...!! It's "so easy" ;-). In *Proceedings of CIKM-Workshop TSA*.

Y. Choi and C. Cardie. 2008. Learning with Compositional Semantics as Structural Inference for Subsentential Sentiment Analysis. In *Proceedings of EMNLP*.

Y. Huang and H. J. Lowe. 2007. A Novel Hybrid Approach to Automated Negation Detection in Clinical Radiology Reports. *JAMIA*, 14.

L. Jia, C. Yu, and W. Meng. 2009. The Effect of Negation on Sentiment Analysis and Retrieval Effectiveness. In *Proceedings of CIKM*.

A. Kennedy and D. Inkpen. 2005. Sentiment Classification of Movie Reviews Using Contextual Valence Shifters. In *Proceedings of FINEXIN*.

A. Kennedy and D. Inkpen. 2006. Sentiment Classification of Movie Reviews Using Contextual Valence Shifters. *Computational Intelligence*, 22.

M. Klenner, S. Petrakis, and A. Fahrni. 2009. Robust Compositional Polarity Classification. In *Proceedings of RANLP*.

L. Ku, T. Huang, and H. Chen. 2009. Using Morphological and Syntactic Structures for Chinese Opinion Analysis. In *Proceedings ACL/IJCNLP*.

J. Liu and S. Seneff. 2009. Review Sentiment Scoring via a Parse-and-Paraphrase Paradigm. In *Proceedings of EMNLP*.

K. Moilanen and S. Pulman. 2007. Sentiment Construction. In *Proceedings of RANLP*.

K. Moilanen and S. Pulman. 2008. The Good, the Bad, and the Unknown. In *Proceedings of ACL/HLT*.

R. Morante and W. Daelemans. 2009. A Metalearning Approach to Processing the Scope of Negation. In *Proceedings of CoNLL*.

R. Morante, A. Liekens, and W. Daelemans. 2008. Learning the Scope of Negation in Biomedical Texts. In *Proceedings of EMNLP*.

V. Ng, S. Dasgupta, and S. M. Niaz Arifin. 2006. Examining the Role of Linguistic Knowledge Sources in the Automatic Identification and Classification of Reviews. In *Proceedings of COLING/ACL*.

B. Pang and L. Lee. 2004. A Sentimental Education: Sentiment Analysis Using Subjectivity Summarization Based on Minimum Cuts. In *Proceedings of ACL*.

B. Pang, L. Lee, and S. Vaithyanathan. 2002. Thumbs up? Sentiment Classification Using Machine Learning Techniques. In *Proceedings of EMNLP*.

L. Polanyi and A. Zaenen. 2004. Context Valence Shifters. In *Proceedings of the AAAI Spring Symposium on Exploring Attitude and Affect in Text*.

M. A. M. Shaikh, H. Prendinger, and M. Ishizuka. 2007. Assessing Sentiment of Text by Semantic Dependency and Contextual Valence Analysis. In *Proceedings of ACII*.

S. Somasundaran and J. Wiebe. 2009. Recognizing Stances in Online Debates. In *Proceedings of ACL/IJCNLP*.

G. Szarvas, V. Vincze, R. Farkas, and J. Csirik. 2008. The BioScope Corpus: Annotation for Negation, Uncertainty and Their Scope in Biomedical Texts. In *Proceedings of BioNLP*.

O. Tsur, D. Davidov, and A. Rappoport. 2010. ICWSM - A Great Catchy Name: Semi-Supervised Recognition of Sarcastic Sentences in Online Product Reviews. In *Proceeding of ICWSM*.

J. Wiebe. 1994. Tracking Point of View in Narrative. *Computational Linguistics*, 20.

T. Wilson, J. Wiebe, and P. Hoffmann. 2005. Recognizing Contextual Polarity in Phrase-level Sentiment Analysis. In *Proceedings of HLT/EMNLP*.

T. Wilson, J. Wiebe, and P. Hoffmann. 2009. Recognizing Contextual Polarity: An Exploration for Phrase-level Analysis. *Computational Linguistics*, 35:3.

T. Zagibalov and J. Carroll. 2008. Automatic Seed Word Selection for Unsupervised Sentiment Classification of Chinese Text. In *Proceedings of COLING*.

Evaluating a Meta-Knowledge Annotation Scheme for Bio-Events

Raheel Nawaz[1] **Paul Thompson**[1,2] **Sophia Ananiadou**[1,2]
[1]School of Computer Science, University of Manchester, UK
[2]National Centre for Text Mining, University of Manchester, UK
E-mail: nawazr@cs.man.ac.uk, paul.thompson@manchester.ac.uk,
sophia.ananiadou@manchester.ac.uk

Abstract

The correct interpretation of biomedical texts by text mining systems requires the recognition of a range of types of high-level information (or *meta-knowledge*) about the text. Examples include expressions of negation and speculation, as well as pragmatic/rhetorical intent (e.g. whether the information expressed represents a hypothesis, generally accepted knowledge, new experimental knowledge, etc.) Although such types of information have previously been annotated at the text-span level (most commonly sentences), annotation at the level of the event is currently quite sparse. In this paper, we focus on the evaluation of the multi-dimensional annotation scheme that we have developed specifically for enriching bio-events with meta-knowledge information. Our annotation scheme is intended to be general enough to allow integration with different types of bio-event annotation, whilst being detailed enough to capture important subtleties in the nature of the meta-knowledge expressed in the text. To our knowledge, our scheme is unique within the field with regards to the diversity of meta-knowledge aspects annotated for each event, whilst the evaluation results have confirmed its feasibility and soundness.

1 Introduction

The ability to recognise high-level information (or meta-knowledge) relating to the interpretation of texts is an important task for text mining systems. There are several types of meta-knowledge that fall under this category. For example, the detection of expressions of speculation and negation is important across all domains, although the way in which these phenomena are expressed may be domain-specific. In scientific texts, it is also important to be able to determine other types of information, such as the author's rhetorical/pragmatic intent (de Waard et al., 2009). This would correspond to whether the information expressed represents a hypothesis, accepted knowledge, new experimental knowledge, etc.

The ability to distinguish between these different types of information can be important for tasks such as building and updating models of biological processes, like pathways (Oda et al., 2008), and curation of biological databases (Ashburner et al., 2000). Central to both of these tasks is the identification of *new knowledge* that can enhance these resources, e.g. to build upon an existing, but incomplete model of a biological process (Lisacek et al., 2005) or to ensure that the database is kept up to date. Any new knowledge added should be supported though evidence, which could include linking hypotheses with experimental findings. It is also important to take into account inconsistencies and contradictions reported in the literature.

The production of annotated corpora can help to train text mining systems to recognise types of meta-knowledge, such as the above. Although a number of such corpora have already been produced, different annotation schemes are required according to the exact domain under consideration, as well as the types of task that will be undertaken by the text mining system.

The work described in this paper is focused on the design and evaluation of the meta-knowledge annotation scheme described in Nawaz et al., (2010). The annotation scheme has been specifically designed to recognise a range of meta-knowledge types for events extracted from biomedical texts (henceforth *bio-events*). The aim is to facilitate the development of more useful systems in the context of various biomedical information extraction (IE) and textual inference (TI) tasks. Although the scheme has been designed

Proceedings of the Workshop on Negation and Speculation in Natural Language Processing, pages 69–77,
Uppsala, July 2010.

for application to existing bio-event corpora, it is intended to be applied to any type of bio-relation corpora, and can easily be tailored for other types of relations/events within the domain.

1.1 Bio-Event Representation of Text

Searching for relevant information in electronic documents is most commonly carried out by entering keywords into a search engine. However, such searches will normally return a huge number of documents, many of which will be irrelevant to the user's needs.

A more promising and efficient way of searching is over *events* that have been extracted from texts through the application of natural language processing methods. An event is a structured representation of a certain piece of information contained within the text, which is usually anchored to a particular word in the text (typically a verb or noun) that is central to the description of the event. Events are often represented by a template-like structure with slots that are filled by the event participants. Each event participant is also assigned a role within the event. These participants can be entities, concepts or even other events. This kind of event representation allows the information contained in a text to be represented as a collection of *nested* events.

A *bio-event* is an event specialised for the biomedical domain. Kim et al. (2008) define a bio-event as a dynamic bio-relation involving one or more participants. These participants can be bio-entities or (other) bio-events, and are each assigned a semantic role/slot like *theme* and *cause* etc. Each bio-event is typically assigned a type/class from a chosen bio-event taxonomy/ontology, e.g., the GENIA Event Ontology (Kim et al., 2008). Similarly, the bio-entities are also assigned types/classes from a chosen bio-term taxonomy/ontology, e.g., the Gene Ontology (Ashburner et al., 2000).

As an example, consider the simple sentence shown in Figure 1**.**

> *The results suggest that the narL gene product activates the nitrate reductase operon.*

Figure 1. A Simple Sentence from a Biomedical Abstract

This sentence contains a single bio-event, anchored to the verb *activates*. Figure 2 shows a typical structured representation of this bio-event.

The fact that the verb is anchored to the verb *activates* allows the event-type of *positive regulation* to be assigned. The event has two slots, i.e. *theme* and *cause* whose labels help to characterise the contribution that the slot filler makes towards the meaning of the event. In this case, the slots are filled by the subject and object of the verb *activates*, both of which correspond to different types of bio-entities (i.e. *operon* and *protein).*

EVENT-TRIGGER: *activates*
EVENT-TYPE: positive_regulation
THEME: *nitrate reductase operon:* operon
CAUSE: *narL gene product*: protein

Figure 2. Typical Structured Representation of the Bio-Event mentioned in Figure 1

IE systems trained to extract bio-events from texts allow users to formulate semantic queries over the extracted events. Such queries can specify semantic restrictions on the events in terms of event types, semantic role labels and named entity types etc. (Miyao et al., 2006), in addition to particular keywords. For example, it would be possible to search only for those texts containing bio-events of type *negative_regulation* where the cause is an entity of type *protein*. Such queries provide a great deal more descriptive power than traditional keyword searches over unstructured documents. Biomedical corpora that have been manually annotated with event level information (e.g., Pyysalo et al., 2007; Kim et al., 2008; Thompson et al., 2009) facilitate the training of systems such as those described above.

Whilst event-based querying has advantages for efficient searching, the extracted events have little practical use if they are not accompanied by meta-knowledge information to aid in their interpretation.

1.2 Existing Meta-knowledge Annotation

Various corpora of biomedical literature (abstracts and/or full papers) have been produced that feature some degree of meta-knowledge annotation. These corpora vary in both the richness of the annotation added, and the type/size of the units at which the meta-knowledge annotation has been performed. Taking the unit of annotation into account, we can distinguish between annotations that apply to continuous text-spans, and annotations that have been performed at the event level.

Text-Span Annotation: Such annotations have mostly been carried out at the sentence level. They normally concentrate on a single aspect (or

dimension) of meta-knowledge, normally either speculation/certainty level, (e.g., Light et al., 2004; Medlock & Briscoe, 2007; Vincze et al., 2008) or general information content/rhetorical intent, e.g., *background, methods*, *results*, *insights*. This latter type of annotation has been attempted both on abstracts, (e.g., McKnight & Srinivasan, 2003; Ruch et al., 2007) and full papers, (e.g. Teufel et al., 1999; Langer et al., 2004; Mizuta & Collier, 2004), with the number of distinct annotation categories varying between 4 and 14.

Despite the availability of these corpora, annotation at the sentence level can often be too granular. In terms of information content, a sentence may describe, for example, both an experimental method and its results. The situation becomes more complicated if a sentence contains an expression of speculation. If this is only marked at the sentence level, there may be confusion about which part(s) of the sentence are affected by the speculative expression.

Certain corpora and associated systems have attempted to address these issues. The BioScope corpus (Vincze et al., 2008) annotates the scopes of negative and speculative keywords, whilst Morante & Daelemans (2009) have trained a system to undertake this task. The scheme described by Wilbur et al. (2006) applies annotation to *fragments* of sentences, which are created on the basis of changes in the meta-knowledge expressed. The scheme consists of multiple annotation dimensions which capture aspects of both certainty and rhetorical/pragmatic intent, amongst other things. Training a system to automatically annotate these dimensions is shown to be highly feasible (Shatkay et al., 2008).

Event-Level Annotation: Explicit annotation of meta-knowledge at the event-level is currently rather minimal within biomedical corpora. Whilst several corpora contain annotations to distinguish positive and negative events (e.g. Sanchez-Graillet & Poesio, 2007; Pyysalo et al., 2007), the annotation of the GENIA Event Corpus (Kim et al., 2008) is slightly more extensive, in that it additionally annotates certainty level. To our knowledge, no existing bio-event corpus has attempted annotation that concerns rhetorical/pragmatic intent.

1.3 The Need for an Event-Centric Meta-Knowledge Annotation Scheme

In comparison to meta-knowledge annotation carried out at the text-span level, the amount of annotation carried out at the event level is quite sparse. The question thus arises as to whether it is possible to use systems trained on text-span annotated corpora to assign meta-knowledge to bio-events, or whether new annotation at the event level is required.

Some corpora seem better suited to this purpose than others – whilst sentence-level annotations are certainly too granular for an event-centric view of the text, sentence fragments, such as those identified by Wilbur et al. (2006), are likely to correspond more closely to the extent of text that describes an event and its slots. Likewise, knowing the scopes of negative and speculative keywords within a sentence may be a useful aid in determining whether they affect the interpretation of a particular event.

However, the information provided in these corpora is still not sufficiently precise for event-level meta-knowledge annotation. Even within a text fragment, there may be several different bio-events, each with slightly different meta-knowledge interpretations. In a similar way, not all events that occur within the scope of a negation or speculation keyword are necessarily affected by it.

Based on these observations, we have developed a meta-knowledge annotation scheme that is specifically tailored to bio-events. Our scheme annotates various different aspects or dimensions of meta-knowledge. A close examination of a large number of relevant bio-events has resulted in a scheme that has some similarities to previously proposed schemes, but has a number of differences that seem especially relevant when dealing with events, e.g. the annotation of the manner of the event. The scheme is intended to be general enough to allow integration with existing bio-event annotation schemes, whilst being detailed enough to capture important subtleties in the nature of the meta-knowledge expressed about the event.

1.4 Lexical Markers of Meta-Knowledge

Most of the existing corpora mentioned above annotate text spans or events with particular categories (e.g. certainty level or general information type) in different meta-knowledge dimensions. However, what they do not normally do is to annotate lexical clues or keywords used to determine the correct values.

A number of previous studies have demonstrated the importance of lexical markers (i.e., words or phrases) that can accompany statements in scientific articles in determining the intended

interpretation of the text (e.g. Hyland, 1996; Rizomilioti 2006). We also performed a similar study (Thompson et al., 2008) although, in contrast to other studies, we took a multi-dimensional approach to the categorisation of such lexical items, acknowledging that several types of important information may be expressed through different words in the same sentence. As an example, let us consider the example sentence in Figure 3.

> *The DNA-binding properties of mutations at positions 849 and 668 may indicate that the catalytic role of these side chains is associated with their interaction with the DNA substrate.*

Figure 3. Example Sentence

The author's pragmatic/rhetorical intent towards the statement that *the catalytic role of these side chains is associated with their interaction with the DNA substrate* is encoded by the word *indicate*, which shows that the statement represents an analysis of the evidence stated at the beginning of the sentence, i.e., that the mutations at positions 849 and 668 have DNA-binding properties. Furthermore, the author's *certainty level* (i.e., their degree of confidence) towards this analysis is shown by the word *may*. Here, the author is uncertain about the validity of their analysis.

Whilst our previous work served to demonstrate that the different aspects of meta-knowledge that can be specified lexically within texts require a multi-dimensional analysis to correctly capture their subtleties, it showed that the presence of particular lexical items is not the only important feature for determining meta-knowledge categories. In particular, their presence does not guarantee that the "expected" interpretation can be assumed (Sándor, 2007). In addition, not all types of meta-knowledge are indicated through explicit markers. Mizuta & Collier (2004) note that *rhetorical zones* may be indicated not only through explicit lexical markers, but also through features such as the main verb in the clause and the position of the sentence within the article or abstract.

For these reasons, we perform annotation on *all* relevant instances, regardless of the presence of lexical markers. This will allow systems to be trained that can learn to determine the correct meta-knowledge category, even when lexical markers are not present. However, due to the proven importance of lexical markers in determining certain meta-knowledge dimensions, our annotation scheme annotates such markers, whenever they are present.

2 Annotation Scheme

The annotation scheme we present here is a slightly modified version of our original meta-knowledge annotation scheme (Nawaz et al., 2010). The modified scheme consists of five meta-knowledge dimensions, each with a set of complete and mutually-exclusive categories, i.e., any given bio-event belongs to exactly one category in each dimension. Our chosen set of annotation dimensions has been motivated by the major information needs of biologists discussed earlier, i.e., the ability to distinguish between different intended interpretations of events.

In order to minimise the annotation burden, the number of possible categories within each dimension has been kept as small as possible, whilst still respecting important distinctions in meta-knowledge that have been observed during our corpus study.

The advantage of using a multi-dimensional scheme is that the interplay between different values of each dimension can reveal both subtle and substantial differences in the types of meta-knowledge expressed in the surrounding text. Therefore, in most cases, the exact rhetorical/pragmatic intent of an event can only be determined by considering a combination of the values of different dimensions. This aspect of our scheme is further discussed in section 3.

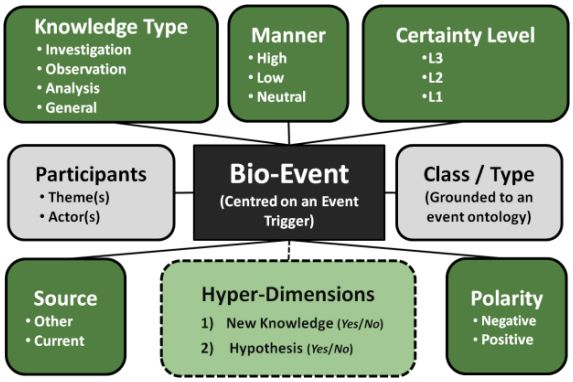

Figure 4. Bio-Event Annotation

Figure 4 provides an overview of the annotation scheme. The boxes with the light-coloured (grey) background correspond to information that is common to most bio-event annotation schemes, i.e., the participants in the event, together with an indication of the class or type of

the event. The boxes with the darker (green) backgrounds correspond to our proposed meta-knowledge annotation dimensions and their possible values. The remainder of this section provides brief details of each annotation dimension.

2.1 Knowledge Type (KT)

This dimension is responsible for capturing the general information content of the event. Whilst less detailed than some of the previously proposed sentence-level schemes, its purpose is to form the basis of distinguishing between the most critical types of rhetorical/pragmatic intent, according to the needs of biologists. Each event is thus classified into one of the following four categories:

Investigation: Enquiries or investigations, which have either already been conducted or are planned for the future, typically marked by lexical clues like *examined*, *investigated* and *studied*, etc.

Observation: Direct observations, often represented by lexical clues like *found*, *observed* and *report*, etc. Simple past tense sentences typically also describe observations. Such events represent experimental knowledge.

Analysis: Inferences, interpretations, speculations or other types of cognitive analysis, typically expressed by lexical clues like *suggest*, *indicate*, *therefore* and *conclude* etc. Such events, if they are interpretations or reliable inferences based on experimental results, can also constitute another type of (indirect) experimental knowledge. Weaker inferences or speculations, however, may be considered as hypotheses which need further proof through experiments.

General: Scientific facts, processes, states or methodology. This is the default category for the knowledge type dimension.

2.2 Certainty Level (CL)

The value of this dimension is almost always indicated through the presence/absence of an explicit lexical marker. In scientific literature, it is normally only applicable to events whose *KT* corresponds either to *Analysis* or *General*. In the case of *Analysis* events, *CL* encodes confidence in the truth of the event, whilst for *General* events, there is a temporal aspect, to account for cases where a particular process is explicitly stated to occur most (but not all) of the time, using a marker such as *normally*, or only occasionally, using a marker like *sometimes.* Events corresponding to direct *Observations* are not open to judgements of certainty, nor are *Investigation* events, which refer to things which have not yet happened or have not been verified.

Regarding the choice of values for the *CL* dimension, there is an ongoing discussion as to whether it is possible to partition the epistemic scale into discrete categories (Rubin, 2007). However, the use of a number of distinct categories is undoubtedly easier for annotation purposes and has been proposed in a number of previous schemes. Although recent work has suggested the use of four or more categories (Shatkay et al., 2008; Thompson et al., 2008), our initial analysis of bio-event corpora has shown that only three levels of certainty seem readily distinguishable for bio-events. This is in line with Hoye (1997), whose analysis of general English showed that there are at least three articulated points on the epistemic scale.

We have chosen to use numerical values for this dimension, in order to reduce potential annotator confusions or biases that may be introduced through the use of labels corresponding to particular lexical markers of each category, such as *probable* or *possible*, and also to account for the fact that slightly different interpretations apply to the different levels, according to whether the event has a *KT* value of *Analysis* or *General*.

L3: No expression of uncertainty or speculation (default category)

L2: High confidence or slight speculation.

L1: Low confidence or considerable speculation; typical lexical markers include *may*, *might* and *perhaps.*

2.3 Source

The source of experimental evidence provides important information for biologists. This is demonstrated by its annotation during the creation of the Gene Ontology (Ashburner et al., 2000) and in the corpus created by Wilbur et al. (2006). The *Source* dimension can also help in distinguishing new experimental knowledge from previously reported knowledge. Our scheme distinguishes two categories, namely:

Other: The event is attributed to a previous study. In this case, explicit clues (citations or phrases like *previous studies* etc.) are normally present.

Current: The event makes an assertion that can be (explicitly or implicitly) attributed to the current study. This is the default category, and is assigned in the absence of explicit lexical or contextual clues.

2.4 Polarity

This dimension identifies negated events. Although certain bio-event corpora are annotated with this information, it is still missing from others. The indication of whether an event is negated is vital, as the interpretation of a negated event instance is completely opposite to the interpretation of a non-negated (positive) instance of the same event.

We define negation as the absence or non-existence of an entity or a process. Negation is typically expressed by the adverbial *not* and the nominal *no*. However, other lexical devices like negative affixals (*un-* and *in-*, etc.), restrictive verbs (*fail*, *lack*, and *unable*, etc.), restrictive nouns (*exception*, etc.), certain adjectives (*independent*, etc.), and certain adverbs (*without*, etc.) can also be used.

2.5 Manner

Events may be accompanied by a word or phrase which provides an indication of the rate, level, strength or intensity of the interaction. We refer to this as the *Manner* of the event. Information regarding manner is absent from the majority of existing bio-event corpora, but yet the presence of such words can be significant in the correct interpretation of the event. Our scheme distinguishes 3 categories of *Manner*, namely:

High: Typically expressed by adverbs and adjectives like *strongly*, *rapidly* and *high*, etc.

Low: Typically expressed by adverbs and adjectives like *weakly*, *slightly* and *slow*, etc.

Neutral: Default category assigned to all events without an explicit indication of manner.

3 Hyper-Dimensions

Determining the pragmatic/rhetorical intent behind an event is not completely possible using any one of our explicitly annotated dimensions. Although the *Knowledge Type* value forms the basis for this, it is not in itself sufficient. However, a defining feature of our annotation scheme is that additional information can be inferred by considering combinations of some of the explicitly annotated dimensions. We refer to this additional information as "latent" or "hyper" dimensions of our scheme. We have identified two such hyper-dimensions.

3.1 New Knowledge

The isolation of events describing new knowledge can be important in certain tasks undertaken by biologists, as explained earlier. Events with the *Knowledge Type* of *Observation* could correspond to new knowledge, but only if they represent observations from the current study, rather than observations cited from elsewhere. In a similar way, an *Analysis* drawn from experimental results in the current study could be treated as new knowledge, but generally only if it represents a straightforward interpretation of results, rather than something more speculative.

Hence, we consider *New Knowledge* to be a hyper-dimension of our scheme. Its value (either *Yes* or *No*) is inferred by considering a combination of the value assignments for the *KT*, *Source* and *CL* dimensions.

Table 1 shows the inference table that can be used to obtain the value for the *New Knowledge* hyper-dimension from the assigned values of the *Source*, *KT* and *CL* dimensions. The symbol 'X' indicates a "don't care condition", meaning that this value does not have any impact on the result.

Source (Annotated)	**KT** (Annotated)	**CL** (Annotated)	**New Knowledge** (Inferred)
Other	X	X	No
X	X	L2	No
X	X	L1	No
Current	Observation	L3	Yes
Current	Analysis	L3	Yes
X	General	X	No
X	Investigation	X	No

Table 1. Inference-Table for New Knowledge Hyper-Dimension

3.2 Hypothesis

A further hyper-dimension of our scheme is *Hypothesis*. The binary value of this hyper-dimension can be inferred by considering the values of *KT* and *CL*. Events with a *KT* value of *Investigation* can always be assumed to be a hypothesis, However, if the KT value is *Analysis*, then only those events with a *CL* value of L1 or L2 (speculative inferences made on the basis of results) should be considered as hypothesis, to be matched with more definite experimental evidence when available. A value of L3 in this instance would normally be classed as new knowledge, as explained in the previous section.

Table 2 shows the inference table that can be used to get the value for the *Hypothesis* hyper-dimension.

KT (Annotated)	CL (Annotated)	Hypothesis (Inferred)
General	X	No
Observation	X	No
Analysis	L3	No
Analysis	L2	Yes
Analysis	L1	Yes
Investigation	X	Yes

Table 2. Inference-Table for Hypothesis Hyper-Dimension

4 Evaluation

The annotation scheme has been evaluated through a small annotation experiment. We randomly choose 70 abstracts from the GENIA Pathway Corpus, which collectively contain over 2600 annotated bio-events. Two of the authors independently annotated these bio-events using a set of annotation guidelines. These guidelines were developed following an analysis of the various bio-event corpora and the output of the initial case study (Nawaz et al., 2010).

The highly favourable results of this experiment further confirmed the feasibility and soundness of the annotation scheme. The remainder of this section discusses the results in more detail.

Dimension	Cohen's Kappa
Knowledge Type	0.9017
Certainty Level	0.9329
Polarity	0.9059
Manner	0.8944
Source	0.9520

Table 3. Inter-Annotator Agreement

4.1 Inter-Annotator Agreement

We have used the familiar measure of Cohen's kappa (Cohen, 1960) for assessing the quality of annotation. Table 3 shows the kappa values for each annotated dimension. The highest value of kappa was achieved for the *Source* dimension, while the *KT* dimension yielded the lowest kappa value. Nevertheless, the kappa scores for all annotation dimensions were in the *good* region (Krippendorff, 1980).

4.2 Category Distribution

Knowledge Type: The most prevalent category found in this dimension was *Observation*, with 45% of all annotated events belonging to this category. Only a small fraction (4%) of these events was represented by an explicit lexical clue (mostly sensory verbs). In most cases the tense, local context (position within the sentence) or global context (position within the document) were found to be important factors.

The second most common category (37% of all annotated events) was *General*. We discovered that most (64%) of the events belonging to this category were processes or states embedded in noun phrases (such as *c-fos expression*). More than a fifth of the *General* events (22%) expressed known scientific facts, whilst a smaller fraction (14%) expressed experimental/scientific methods (such as *stimulation* and *incubation* etc.). Explicit lexical clues were found only for facts, and even then in only 1% of cases.

Analysis was the third most common category, comprising 16% of all annotated events. Of the events belonging to this category, 44% were deductions (*CL=L1*), whilst the remaining 54% were hedged interpretations (*CL=L2/L3*). All *Analysis* events were marked with explicit lexical clues.

The least common category was *Investigation* (1.5% of all annotated events). All *Investigation* events were marked with explicit lexical clues.

Certainty Level: *L3* was found to be the most prevalent category, corresponding to 93% of all events. The categories *L2* and *L1* occurred with frequencies of 4.3% and 2.5%, respectively. The relative scarcity of speculative sentences in scientific literature is a well documented phenomenon (Thompson et al., 2008; Vincze et al., 2008). Vincze et al. (2008) found that less than 18% of sentences occurring in biomedical abstracts are speculative. Similarly, we found that around 20% of corpus events belong to speculative sentences. Since speculative sentences contain non-speculative events as well, the frequency of speculative events is expected to be much less than the frequency of speculative sentences. In accordance with this hypothesis, we found that only 7% of corpus events were expressed with some degree of speculation. We also found that almost all speculated events had explicit lexical clues.

Polarity: Our event-centric view of negation showed just above 3% of the events to be negated. Similarly to speculation, the expected fre-

quency of negated events is lower than the frequency of negated sentences. Another reason for finding fewer negated events is the fact that, in contrast to previous schemes, we draw a distinction between events that are negated and events expressed with *Low* manner. For example, certain words like *limited* and *barely* are often considered as negation clues. However, we consider them as clues for *Low* manner. In all cases, negation was expressed through explicit lexical clues.

Manner: Whilst only a small fraction (4%) of events contains an indication of *Manner*, we found that where present, manner conveys vital information about the event. Our results also revealed that indications of *High* manner are three times more frequent than the indications of *Low* manner. We also noted that both *High* and *Low* manners were always indicated through the use of explicit clues.

Source: Most (99%) of the events were found to be of the *Current* category. This is to be expected, as authors tend to focus on current work in within abstracts. It is envisaged, however, that this dimension will be more useful for analyzing full papers.

Hyper-dimensions: Using the inference tables shown in section 3, we calculated that almost 57% of the events represent *New Knowledge*, and just above 8% represent *Hypotheses*.

5 Conclusion and Future Work

We have evaluated a slightly modified version of our meta-knowledge annotation scheme for bio-events, first presented in Nawaz et al. (2010). The scheme captures key information regarding the correct interpretation of bio-events, which is not currently annotated in existing bio-event corpora, but which we have shown to be critical in a number of text mining tasks undertaken by biologists. The evaluation results have shown high inter-annotator agreement and a sufficient number of annotations along each category in every dimension. These results have served to confirm the feasibility and soundness of the annotation scheme, and provide promising prospects for its application to existing and new bio-event corpora.

We are currently working on a large scale annotation effort, involving multiple independent annotators. Although our main objective is to enrich the entire GENIA event corpus with meta-knowledge information, we also plan to create a small corpus of full papers enriched with bio-event and meta-knowledge annotations.

Acknowledgments

The work described in this paper has been funded by the Biotechnology and Biological Sciences Research Council through grant numbers BBS/B/13640, BB/F006039/1 (ONDEX)

References

M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin and G. Sherlock. 2000. Gene ontology: tool for the unification of biology. *Nature Genetics* 25:25-29.

J. Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement* 20: 37–46.

A. de Waard, B. Shum, A. Carusi, J. Park, M. Samwald and Á. Sándor. 2009. Hypotheses, Evidence and Relationships: The HypER Approach for Representing Scientific Knowledge Claims. In *Proceedings of the Workshop on Semantic Web Applications in Scientific Discourse*. Available at: http://oro.open.ac.uk/18563/

L. Hoye. 1997. *Adverbs and Modality in English*. London & New York: Longman

K. Hyland. 1996. Talking to the Academy: Forms of Hedging in Science Research Articles. *Written Communication* 13(2):251-281.

K. Hyland. 2005. *Metadiscourse: Exploring Interaction in Writing*. London: Continuum

J. Kim, T. Ohta and J. Tsujii. 2008. Corpus annotation for mining biomedical events from literature. *BMC Bioinformatics* 9:10

K. Krippendorff. 1980. *Content Analysis: An Introduction to Its Methodology*. Beverly Hills: Sage Publications

H. Langer, H. Lungen and P. S. Bayerl. 2004. Text type structure and logical document structure. In *Proceedings of the ACL Workshop on Discourse Annotation*, pages 49-56

M. Light, X. T. Qui and P. Srinivasan. 2004. The language of bioscience: Facts, speculations, and statements in between. In *Proceedings of the BioLink 2004 Workshop on Linking Biological Literature, Ontologies and Databases: Tools for Users*, pages 17-24.

F. Lisacek, C. Chichester, A. Kaplan and A. Sandor. 2005. Discovering Paradigm Shift Patterns in Biomedical Abstracts: Application to Neurodegenerative Diseases. In *Proceedings of SMBM 2005*, pages 212-217

L. McKnight and P. Srinivasan. 2003. Categorization of sentence types in medical abstracts. In *Proceedings of the 2003 Annual Symposium of AMIA*, pages 440-444.

B. Medlock and T. Briscoe. 2007. Weakly supervised learning for hedge classification in scientific literature. In *Proceedings of ACL 2007*, pages 992- 999.

Y. Miyao, T. Ohta, K. Masuda, Y. Tsuruoka, K. Yoshida, T. Ninomiya and J. Tsujii. 2006. Semantic Retrieval for the Accurate Identification of Relational Concepts in Massive Textbases. In *Proceedings of COLING-ACL 2006*, pages 1017-1024.

Y. Mizuta and N. Collier. 2004. Zone identification in biology articles as a basis for information extraction. In *Proceedings of the joint NLPBA/BioNLP Workshop on Natural Language for Biomedical Applications*, pages 119-125.

R. Morante and W. Daelemans. 2009. A metalearning approach to processing the scope of negation. In *Proceedings of CoNLL 2009*, pages 21-29.

R. Nawaz, P. Thompson, J. McNaught and S. Ananiadou. 2010. Meta-Knowledge Annotation of Bio-Events. In *Proceedings of LREC 2010*, pages 2498-2507.

K. Oda, J. Kim, T. Ohta, D. Okanohara, T. Matsuzaki, Y. Tateisi and J. Tsujii. 2008. New challenges for text mining: mapping between text and manually curated pathways. *BMC Bioinformatics* 9(Suppl 3): S5.

S. Pyysalo, F. Ginter, J. Heimonen, J. Bjorne, J. Boberg, J. Jarvinen and T. Salakoski. 2007. BioInfer: a corpus for information extraction in the biomedical domain. *BMC Bioinformatics* 8:50.

V. Rizomilioti. 2006. "Exploring Epistemic Modality in Academic Discourse Using Corpora." *Information Technology in Languages for Specific Purposes* 7, pages 53-71

V. L. Rubin. 2007. Stating with certainty or stating with doubt: Intercoder reliability results for manual annotation of epistemically modalized statements. In *Proceedings of NAACL-HLT 2007, Companion Volume***,** pages 141-144.

P. Ruch, C. Boyer, C. Chichester, I. Tbahriti, A. Geissbühler, P. Fabry, J. Gobeill, V. Pillet, D. Rebholz-Schuhmann and C. Lovis. 2007. Using argumentation to extract key sentences from biomedical abstracts. *International Journal of Medical Informatics* 76(2-3):195-200.

O. Sanchez-Graillet and M. Poesio. 2007. Negation of protein-protein interactions: analysis and extraction. *Bioinformatics* 23(13):i424-i432

Á. Sándor. 2007. Modeling metadiscourse conveying the author's rhetorical strategy in biomedical research abstracts. *Revue Française de Linguistique Appliquée* 200(2):97-109.

H. Shatkay, F. Pan, A. Rzhetsky and W. J. Wilbur. 2008. Multi-dimensional classification of biomedical text: toward automated, practical provision of high-utility text to diverse users. *Bioinformatics* 24(18): 2086-2093.

S. Teufel, J. Carletta and M. Moens. 1999. An annotation scheme for discourse-level argumentation in research articles. In *Proceedings of EACL 1999*, pages 110-117.

S. Teufel, A. Siddharthan and C. Batchelor. 2009. Towards discipline-independent argumentative zoning: Evidence from chemistry and computational linguistics. In *Proceedings of EMNLP-09*, pages 1493-1502

P. Thompson, S. Iqbal, J. McNaught and S. Ananiadou. 2009. Construction of an annotated corpus to support biomedical information extraction. *BMC Bioinformatics* 10: 349.

P. Thompson, G. Venturi, J. McNaught, S. Montemagni and S. Ananiadou. 2008. Categorising Modality in Biomedical Texts. In *Proceedings of the LREC 2008 Workshop on Building and Evaluating Resources for Biomedical Text Mining*, pages 27-34.

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.

W. J. Wilbur, A. Rzhetsky and H. Shatkay. 2006. New directions in biomedical text annotations: definitions, guidelines and corpus construction. *BMC Bioinformatics 7*: 356.

Using SVMs with the Command Relation Features to Identify Negated Events in Biomedical Literature

Farzaneh Sarafraz
School of Computer Science
University of Manchester
Manchester, United Kingdom
sarafraf@cs.man.ac.uk

Goran Nenadic
School of Computer Science
University of Manchester
Manchester, United Kingdom
g.nenadic@manchester.ac.uk

Abstract

In this paper we explore the identification of negated molecular events (e.g. protein binding, gene expressions, regulation, etc.) in biomedical research abstracts. We construe the problem as a classification task and apply a machine learning (ML) approach that uses lexical, syntactic, and semantic features associated with sentences that represent events. Lexical features include negation cues, whereas syntactic features are engineered from constituency parse trees and the *command* relation between constituents. Semantic features include event type and participants. We also consider a rule-based approach that uses only the command relation. On a test dataset, the ML approach showed significantly better results (51% F-measure) compared to the command-based rules (35-42% F-measure). Training a separate classifier for each event class proved to be useful, as the micro-averaged F-score improved to 63% (with 88% precision), demonstrating the potential of task-specific ML approaches to negation detection.

1 Introduction

With almost 2000 new papers published every day, biomedical knowledge is mainly communicated through a growing body of research papers. As the amount of textual information increases, the need for sophisticated information extraction (IE) methods are becoming more than evident. IE methods rely on a range of language processing methods such as named entity recognition and parsing to extract the required information in a more structured form which can be used for knowledge exploration and hypothesis generation (Donaldson et al. 2003; Natarajan et al. 2006).

Given the large number of publications, the identification of conflicting or contradicting facts is critical for systematic mining of biomedical literature and knowledge consolidation. Detection of negations is of particular importance for IE methods, as it often can hugely affect the quality of the extracted information. For example, when mining molecular events, a key piece of information is whether the text states that the two proteins *are* or *are not* interacting, or that a given gene *is* or *is not* expressed. In recent years, several challenges and shared tasks have included the extraction of negations, typically as part of other tasks (e.g. the BioNLP'09 Shared Task 3 (Kim et al. 2009)).

Several systems and methods have aimed to handle negation detection in order to improve the quality of extracted information (Hakenberg et al. 2009; Morante and Daelemans 2009). Prior research on this topic has primarily focused on finding negated concepts by negation cues and scopes. These concepts are usually represented by a set of predefined terms, and negation detection typically aims to determine whether a term falls within the scope of a negation cue.

In this paper we address the task of identification of negated events. We present a machine learning (ML) method that combines a set of features mainly engineered from a sentence parse tree with lexical cues. More specifically, parse-based features use the notion of the *command* relation that models the scope affected by an element (Langacker, 1969). We use molecular events as a case study and experiment on the BioNLP'09 data, which comprises a gold-standard corpus of research abstracts manually annotated for events and negations (Kim et al. 2009). The evaluation shows that, by using the proposed approach, negated events can be identified with precision of 88% and recall of 49% (63% F-measure). We compare these results with two rule-based approaches that achieved the maximum F-measure of 42%.

Proceedings of the Workshop on Negation and Speculation in Natural Language Processing, pages 78–85,
Uppsala, July 2010.

The rest of this paper is organised as follows. Section 2 summarises and reviews previous research on negation extraction. Section 3 defines the problem and introduces the data used for the case study. Section 4 focuses on the ML-based methodology for extracting negated events. The final sections contain the results and discussions.

2 Related Work

There have been numerous contemplations of the concept of negation (Lawler, 2010), but no general agreement so far exists on its definition, form, and function. We adopt here a definition of negation as given by Cambridge Encyclopedia of Language Sciences: "Negation is a comparison between a 'real' situation lacking some element and an 'imaginal' situation that does not lack it". The imaginal situation is **affirmative** compared with the **negative** real situation. The element whose polarity differs between the two situations is the negation **target**.

Negations in natural language can be expressed by syntactically negative expressions, i.e. with the use of negating words such as *no*, *not*, *never*, etc. The word or phrase that makes the sentence wholly or partially negative is the negation **cue** and the part of the sentence that is affected by the negation cue and has become negative is the negation **scope**.

We briefly review two classes of approaches to detect negations: those aiming at negated concepts and those targeting negated events.

2.1 Detecting Negated Concepts and Phrases

There have been a number of approaches suggested for detection of negated targets and scopes. Most of them rely on task-specific, handcrafted rules of various complexities. They differ in the size and composition of the list of negation cues, and the way to utilise such a list. Some methods use parse trees, whilst others use results of shallow parsing.

Rule-based methods range from simple co-occurrence based approaches to patterns that rely on shallow parsing. The 'bag-of-words' approach, looking for proximate co-occurrences of negation cues and terms in the same sentence, is probably the simplest method for finding negations, and is used by many as a baseline method.

Many approaches have targeted the clinical and biomedical domains. NegEx (Chapman et al. 2001), for example, uses two generic regular expressions that are triggered by negation phrases such as:

<negation cue> * <target term>
<target term> * <negation cue>

where the asterisk (*) represents a string of up to five tokens. Target terms represent domain concepts that are terms from the Unified Medical Language System (UMLS[1]). The cue set comprises 272 clinically-specific negation cues, including those such as *denial of* or *absence of*. Although simple, the proposed approach showed good results on clinical data (78% sensitivity (recall), 84% precision, and 94% specificity).

In addition to concepts that are explicitly negated by negation phrases, Patrick et al. (2006) further consider so-called pre-coordinated negative terms (e.g. *headache*) that have been collected from SNOMED CT[2] medical terminology. Similarly, NegFinder uses hand-crafted rules to detect negated UMLS terms, including simple conjunctive and disjunctive statements (Mutalik et al. 2001). They used a list of 60 negation cues. Tolentino et al. (2006), however, show that using rules on a small set of only five negation cues (*no*, *neither/nor*, *ruled out*, *denies*, *without*) can still be reasonably successful in detecting negations in medical reports (F-score 91%).

Huang and Lowe (2007) introduced a negation grammar that used regular expressions and dependency parse trees to identify negation cues and their scope in the sentence. They applied the rules to a set of radiology reports and reported a precision of 99% and a recall of 92%.

Not many efforts have been reported on using machine learning to detect patterns in sentences that contain negative expressions. Still, Morante and Daelemans (2009), for example, used various classifiers (Memory-based Learners, Support Vector Machines, and Conditional Random Fields) to detect negation cues and their scope. An extensive list of features included the token's stem and part-of-speech, as well as those of the neighbouring tokens. Separate classifiers were used for detecting negation cues and negation scopes. The method was applied to clinical text, biomedical abstracts, and biomedical papers with F-scores of 80%, 77%, and 68% respectively.

2.2 Detecting Negated Events

Several approaches have recently been suggested for the extraction of negated events, particularly

[1] http://www.nlm.nih.gov/research/umls/
[2] http://www.snomed.org

in the biomedical domain. Events are typically represented via *participants* (biomedical entities that take part in an event) and event *triggers* (tokens that indicate presence of the event). Van Landeghem et al. (2008) used a rule-based approach based on token distances in sentence and lexical information in event triggers to detect negated molecular events. Kilicoglu and Bergler (2009), Hakenberg et al. (2009), and Sanchez (2007) used a number of heuristic rules concerning the type of the negation cue and the type of the dependency relation to detect negated molecular events described in text. For example, a rule can state that if the negation cue is "lack" or "absence", then the trigger has to be in the prepositional phrase of the cue; or that if the cue is "unable" or "fail", then the trigger has to be in the clausal complement of the cue (Kilicoglu and Bergler 2009). As expected, such approaches suffer from lower recall.

MacKinlay et al. (2009), on the other hand, use ML, assigning a vector of complex deep parse features (including syntactic predicates to capture negation scopes, conjunctions and semantically negated verbs) to every event trigger. The system achieved an F-score of 36% on the same dataset as used in this paper.

We note that the methods mentioned above mainly focus on finding negated triggers in order to detect negated events. In this paper we explore not only negation of triggers but also phrases in which participants are negated (consider, for example, "SLP-76" in the sentence "*In contrast, Grb2 can be coimmunoprecipitated with Sos1 and Sos2 but not with SLP-76*.")

3 Molecular Events

As a case study, we look at identification of negated molecular events. In general, molecular events include various types of reactions that affect genes and protein molecules. Each event is of a particular *type* (e.g. binding, phosphorylation, regulation, etc.). Depending on the type, each event may have one or more participating proteins (sometimes referred to as *themes*). Regulatory events are particularly complex, as they can have a *cause* (a protein or another event) in addition to a theme, which can be either a protein or another event. Table 1 shows examples of five events, where participants are biomedical entities (events 1-3) or other events (events 4 and 5). Note that a sentence can express more than one molecular event.

Identification of molecular events in the literature is a challenging IE task (Kim et al. 2009; Sarafraz et al. 2009). For the task of identifying negated events, we assume that events have already been identified in text. Each event is represented by its type, a textual trigger, and one or more participants or causes (see Table 1). Since the participants of different event types can vary in both their number and type, we consider three classes of events to support our analysis (see Section 5):

- Class I comprises events with exactly one entity theme (e.g. transcription, protein catabolism, localization, gene expression, phosphorylation).
- Class II events include binding events only, which have one or more entity participants.
- Class III contains regulation events, which have exactly one theme and possibly one cause. However, the theme and the cause can be entities or events of any type.

The corpus used in this study is provided by the BioNLP'09 challenge (Kim et al. 2009). It contains two sets of biomedical abstracts: a "training" set (containing 800 abstracts used for training and analysis purposes) and a "development" set (containing 150 abstracts used for testing purposes only). Both document sets are manually annotated with information about entity mentions (e.g. genes and proteins). Sentences that report molecular events are further annotated with the corresponding event type, textual trigger and participants. In total, nine event types are

"The effect of this synergism was perceptible at the level of induction of the IL-2 gene."				
Event	**Trigger**	**Type**	**Participant (theme)**	**Cause**
Event 1	"induction"	Gene expression	IL-2	

"Overexpression of full-length ALG-4 induced transcription of FasL and, consequently, apoptosis."				
Event	**Trigger**	**Type**	**Participant (theme)**	**Cause**
Event 2	"transcription"	Transcription	FasL	
Event 3	"Overexpression"	Gene expression	ALG-4	
Event 4	"Overexpression"	Positive regulation	Event 3	
Event 5	"induced"	Positive regulation	Event 2	Event 4

Table 1: Examples of how molecular events described in text are characterised.

considered (gene expression, transcription, protein catabolism, localization, phosphorylation, binding, regulation, positive regulation, and negative regulation). In addition, every event has been tagged as either affirmative (reporting a specific interaction) or negative (reporting that a specific interaction has not been observed).

Table 2 provides an overview of the two BioNLP'09 datasets. We note that only around 6% of events are negated.

Event class	Training data		Development data	
	total	negated	total	negated
Class I	2,858	131	559	26
Class II	887	44	249	15
Class III	4,870	440	987	66
Total	**9,685**	**615**	**1,795**	**107**

Table 2: Overview of the total number of events and negated event annotations in the two datasets.

4 Methodology

We consider two approaches to extract negated events. We first discuss a rule-based approach that uses constituency parse trees and the *command* relation to identify negated events. Then, we introduce a ML method that combines lexical, syntactic and semantic features to identify negated events. Note that in all cases, input sentences have been pre-annotated for entity mentions, event triggers, types, and participants.

4.1 Negation Detection Using the *Command* Relation Rules

The question of which parts of a syntactic structure affect the other parts has been extensively investigated. Langacker (1969) introduced the concept of *command* to determine the scope within a sentence affected by an element. More precisely, if *a* and *b* are nodes in the constituency parse tree of a sentence, then *a* X-commands *b* iff the lowest ancestor of *a* with label X is also an ancestor of *b*. Note that the command relation is not symmetrical. Langacker observed that when *a* S-commands *b*, then *a* affects the scope containing *b*. For simplicity, we say "command" when we mean S-command.

To determine whether token *a* commands token *b*, given the parse tree of a sentence, we use a simple algorithm introduced by McCawley (1993): trace up the branches of the constituency parse tree from *a* until you hit a node that is labelled X. If *b* is reachable by tracing down the branches of the tree from that node, then *a* X-commands *b*; otherwise, it does not.

We hypothesise that if a negation cue commands an event trigger *or* participant, then the associated event is negated.

4.2 Negation Detection Using Machine Learning on Parse Tree Features

Given a sentence that describes an event, we further construe the negation detection problem as a classification task: the aim is to classify the event as affirmative or negative. We explore both a single SVM (support vector machine) classifier for all events and three separate SVMs for each of the event classes. The following features have been engineered from an event-representing sentence:

1. Event type (one of the nine types as defined in BioNLP'09);
2. Whether the sentence contains a negation cue from the cue list;
3. The negation cue itself (if present);
4. The part-of-speech (POS) tag of the negation cue;
5. The POS tag of the event trigger;
6. The POS tag of the participants of the event. If the participant is another event, the POS tag of the trigger of that event is used;
7. The parse node type of the lowest common ancestor of the trigger and the cue (i.e. the type of the smallest phrase that contains both the trigger and the cue, e.g. S, VP, PP, etc.);
8. Whether or not the negation cue commands any of the participants; nested events (for Class III) are treated as above (i.e. as being represented by their triggers);
9. Whether or not the negation cue commands the trigger;
10. The parse-tree distance between the event trigger and the negation cue.

We use a default value (null) where none of the other values apply (e.g. when there is no cue in feature 3, 4, 7). These features have been used to train four SVMs on the training dataset: one modelled all events together, and the others modelled the three event classes separately.

5 Results

All the results refer to the methods applied on the development dataset (see Table 2). If the negation detection task is regarded as an information extraction task of finding positive instances (i.e.

negated events), then precision, recall, and F-score would be appropriate measures. If we consider the classification aspect of the task, specificity is more appropriate if true negative hits are considered as valuable as true positive ones. We therefore use the following metrics to evaluate the two methods:

$$Precision = \frac{TP}{TP+FP}$$

$$Recall = Sensitivity = \frac{TP}{TP+FN}$$

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

$$Specificity = \frac{TN}{TN+FP}$$

where TP denotes the number of true positives (the number of correctly identified negated events), FN is the number of false negatives (the number of negated events that have been reported as affirmative), with TN and FP defined accordingly.

Two sets of negation cues were used in order to compare their influence. A smaller set was derived from related work, whereas additional cues were semi-automatically extracted by exploring the training data. The small negation cue set contains 14 words[3], whereas the larger negation cue set contains 32 words[4]. As expected, the larger set resulted in increased recall, but decreased precision. However, the effects on the F-score were typically not significant. The results are only shown using the larger cue set.

The texts were processed using the GENIA tagger (Tsuruoka and Tsujii 2005).We used constituency parse trees automatically produced by two different constituency parsers reported in (McClosky et al. 2006) and (Bikel 2004). No major differences were observed in the results using the two parsers. The data shown in the results are produced by the former.

5.1 Baseline Results

Our baseline method relies on an implementation of the NegEx algorithm as explained in Section 2.1. Event triggers were used as negation targets for the algorithm. An event is then considered to be negated if the trigger is negated; otherwise it is affirmative. The results (see Table 3) are substantially lower than those reported for NegEx on clinical data (specificity of 94% and sensitivity of 78%). For comparison, the table also provides an even simpler baseline approach that tags as negated any event whose associated sentence contains any negation cue word.

Approach	P	R	F1	Spec.
any negation cue present	20%	**78%**	32%	81%
NegEx	**36%**	37%	**36%**	**93%**

Table 3: Baseline results.
(NegEx and a 'bag-of-words' approach)

5.2 Rules Based on the *Command* Relation

Table 4 shows the results of applying the S-command relation rule for negation detection. We experimented with three possible approaches: an event is considered negated if

- the negation cue commands any event participant in the parse tree;
- the negation cue commands the event trigger in the tree;
- the negation cue commands both.

Approach	P	R	F1	Spec.
negation cue commands any participant	23%	**76%**	**35%**	84%
negation cue commands trigger	23%	68%	34%	85%
negation cue commands both	23%	68%	35%	**86%**

Table 4: Performance when only the S-command relation is used.

Compared with the baseline methods, the rules based on the command relation did not improve the performance. While precision was low (23%), recall was high (around 70%), indicating that in the majority of cases there is an S-command relation in particular with the participants (the highest recall). We also note a significant drop in specificity, as many affirmative events have triggers/participants S-commanded by a negation cue (not "linked" to a given event).

5.3 Machine Learning Results

All SVM classifiers have been trained on the training dataset using a Python implementation of SVM Light using the linear kernel and the default parameters (Joachims 1999). Table 5 shows the results of the single SVM classifier that has been trained for all three event classes together (applied on the development data).

[3] Negation cues in this set include: *no, not, none, negative, without, absence, fail, fails, failed, failure, cannot, lack, lacking, lacked.*

[4] Negation cues in this set include the smaller set and 18 task-specific words: *inactive, neither, nor, inhibit, unable, blocks, blocking, preventing, prevents, absent, never, unaffected, unchanged, impaired, little, independent, except,* and *exception.*

Compared to previous methods, there was significant improvement in precision, while recall was relatively low. Still, the overall F-measure was significantly better compared with the rule-based methods (51% vs. 35%).

Feature set	P	R	F1	Spec.
Features 1-7	43%	8%	14%	99.2%
Features 1-8	73%	19%	30%	**99.3%**
Features 1-9	71%	38%	49%	99.2%
Features 1-10	**76%**	**38%**	**51%**	99.2%

Table 5: The results of the single SVM classifier. Features 1-7 are lexical and POS tag-based features. Feature 8 models whether the cue S-commands any of the participants. Feature 9 is related to the cue S-commanding the trigger. Feature 10 is the parse-tree distance between the cue and trigger.

We first experimented with the effect of different types of feature on the quality of the negation prediction. Table 5 shows the results of the first classifier with an incremental addition of lexical features, parse tree-related features, and finally a combination of those with the command relation between the negation cue and event trigger and participants. It is worth noting that both precision and recall improved as more features are added.

We also separately trained classifiers on the three classes of events (see Table 6). This further increased the performance: compared with the results of the single classifier, the F1 micro-average improved from 51% to 63%, with similar gains for both precision and recall.

Event class	P	R	F1	Spec.
Class I (559 events)	94%	**65%**	**77%**	99.8%
Class II (249 events)	**100%**	33%	50%	**100%**
Class III (987 events)	81%	44%	57%	99.2%
Micro Average (1,795 events)	88%	49%	63%	99.4%
Macro Average (3 classes)	92%	47%	62%	99.7%

Table 6: The results of the separate classifiers on different classes using common features.

6 Discussion

As expected, approaches that focus only on event triggers and their surface distances from negation cues proved inadequate for biomedical scientific articles. Low recall was mainly caused by many event triggers being too far from the negation cue to be detected as within the scope.

Furthermore, compared to clinical notes, for example, sentences that describe molecular events are significantly more complex. For example, the event-describing sentences in the training data have on average 2.6 event triggers. The number of events per sentence is even higher, as the same trigger can indicate multiple events, sometimes with opposite polarities. Consider for example the sentence

> *"We also demonstrate that the IKK complex, but not p90 (rsk), is responsible for the in vivo phosphorylation of I-kappa-B-alpha mediated by the co-activation of PKC and calcineurin."*

Here, the trigger (phosphorylation) is linked with one affirmative and one negative regulatory event by two different molecules, hence triggering two events of opposite polarities.

These findings, together with previous work, suggested that for any method to effectively detect negations, it should be able to link the negation cue to the specific token, event trigger or entity name in question. Therefore, more complex models are needed to capture the specific structure of the sentence as well as the composition of the interaction and the arrangement of its trigger and participants.

By combining several feature types (lexical, syntactic and semantic), the machine learning approach proved to provide significantly better results. In the incremental feature addition exploration process, adding the cue-commands-participant feature had the greatest effect on the F-score, suggesting the significance of treating event participants. We note, however, that many of the previous attempts focus on event triggers only, although participants do play an important role in the detection of negations in biomedical events and thus should be used as negation targets instead of or in addition to triggers. It is interesting that adding parse-tree distance between the trigger and negation cue improves precision by 5%.

Differences in event classes (in the number and type of participants) proved to be important. Significant improvement in performance was observed when individual classifiers were trained for the three event classes, suggesting that events with different numbers or types of participants are expressed differently in text, at least when negations are considered. Class I events are the simplest (one participant only), so it was expected that negated events in this class would be

the easiest to detect (F-score of 77%). Class II negated events (which can have multiple participants), demonstrated the lowest recall (33%). A likely reason is that the feature set used is not suitable for multi-participant events: for example, feature 8 focuses on the negation cue commanding *any* of the participants, and not *all* of them. It is surprising that negated regulation events (Class III) were not the most difficult to identify, given their complexity.

We applied the negation detection on the type, trigger and participants of pre-identified events in order to explore the complexity of negations, unaffected by automatic named entity recognition, event trigger detection, participant identification, etc. As these steps are typically performed before further characterisation of events, this assumption is not superficial and such information can be used as input to the negation detection module. MacKinlay et al. (2009) also used gold annotations as input for negation detection, and reported precision, recall, and F-score of 68%, 24%, and 36% respectively on the same dataset (compared to 88%, 49% and 63% in our case). The best performing negation detection approach in the BioNLP'09 shared task reported recall of up to 15%, but with overall event detection sensitivity of 33% (Kilicoglu and Bergler 2009) on a 'test' dataset (different from that used in this study). This makes it difficult to directly compare their results to our work, but we can still provide some rough estimates: had all events been correctly identified, their negation detection approach could have reached 45% recall (compared to 49% in our case). With precision of around 50%, their projected F-score, again assuming perfect event identification, could have been in the region of 50% (compared to 63% in our case).

The experiments with rules that were based on the command relations have proven to be generic, providing very high recall (~70%) but with poor precision. Although only the results with S-command relations have been reported here (see Table 4), we examined other types of command relation, namely NP-, PP-, SBAR-, and VP-command. The only variation able to improve prediction accuracy was whether the cue VP-commands any of the participants, with an F-score of 42%, which is higher than the results achieved by the S-command (F-score of 35%). The S-command relation was used in the SVM modules as VP-command did not make the results significantly better.

One of the issues we faced was the management of multi-token and sub-token entities and triggers (e.g. *alpha B1* and *alpha B2* in "alpha B1/alpha B2 ratio", which will be typically tokenised as "alpha", "B1/alpha", and "B2"). In our approach, we considered all the entities that are either multi-token or sub-token. However, if we assign participants that are *both* multi-token and sub-token simultaneously to events and extract similar features for the classifier from them as from simple entities, the F-score is reduced by about 2%. It would be probably better to assign a new category to those participants and add a new value for them specifically in every feature.

7 Conclusions

Given the number of published articles, detection of negations is of particular importance for biomedical IE. Here we explored the identification of negated molecular events, given their triggers (to characterise event type) and participants. We considered two approaches:[5] a rule-based approach using constituency parse trees and the command relation to identify negation cues and scopes, and a machine learning method that combines a set of lexical, syntactic and semantic features engineered from the associated sentence. When compared with a regular-expression-based baseline method (NegEx-like), the proposed ML method achieved significantly better results: 63% F-score with 88% precision. The best results were obtained when separate classifiers were trained for each of the three event classes, as differences between them (in the number and type of participants) proved to be important.

The results presented here were obtained by using the 'gold' event annotations as the input. It would be interesting to explore the impact of typically noisy automatic event extraction on negation identification. Furthermore, an immediate future step would be to explore class-specific features (e.g. type of theme and cause for Class III events, and whether the cue S-commands all participants for Class II events). In addition, in the current approach we used constituency parse trees. Our previous attempts to identify molecular events (Sarafraz et al. 2009) as well as those discussed in Section 2 use dependency parse trees. A topic open for future research will be to combine information from both dependency and constituency parse trees as features for detecting negated events.

[5] Available at `http://bit.ly/bzBaUX`

Acknowledgments
We are grateful to the organisers of BioNLP'09 for providing the annotated data.

References

Daniel Bikel. 2004. A Distributional Analysis of a Lexicalized Statistical Parsing. *Proc. Conference on Empirical Methods in Natural Language.*

Wendy Chapman. 2001. A Simple Algorithm for Identifying Negated Findings and Diseases in Discharge Summaries. *Journal of Biomedical Informatics*, 34(5):301-310.

Ian Donaldson, Martin, J., de Bruijn, B., Wolting, C., Lay, V., Tuekam, B., Zhang, S., Baskin, B., Bader, G. D., Michalickova, K., Pawson, T. and Hogue, C. W. 2003. PreBIND and Textomy--mining the biomedical literature for protein-protein interactions using a support vector machine. *BMC Bioinf.* 4: 11.

Jörg Hakenberg, Illés Solt, Domonkos Tikk, Luis Tari, Astrid Rheinländer, Quang L. Ngyuen, Graciela Gonzalez and Ulf Leser. 2009. Molecular event extraction from link grammar parse trees. *BioNLP'09: Proceedings of the Workshop on BioNLP*. 86-94.

Yang Huang and Henry J. Lowe. 2007. A Novel Hybrid Approach to Automated Negation Detection in Clinical Radiology Reports. *Journal of the American Medical Informatics Association*, 14(3):304-311.

Thorsten Joachims. 1999. *Making large-Scale SVM Learning Practical.* Advances in Kernel Methods - Support Vector Learning, B. Schölkopf and C. Burges and A. Smola (ed.) MIT-Press, MA.

Halil Kilicoglu, and Sabine Bergler. 2009. Syntactic dependency based heuristics for biological event extraction. *BioNLP'09: Proceedings of the Workshop on BioNLP*. 119-127.

Jin-Dong Kim, Tomoko Ohta, Sampo Pyysalo, Yoshinobu Kano, Jun'ichi Tsujii. 2009. Overview of BioNLP'09 shared task on event extraction. *BioNLP'09: Proceedings of the Workshop on BioNLP*. 1-9.

Sofie Van Landeghem, Yvan Saeys, Bernard De Baets and Yves Van de Peer. 2008. Extracting Protein-Protein Interactions from Text using Rich Feature Vectors and Feature Selection. *Proceedings of the Third International Symposium on Semantic Mining in Biomedicine.* 77-84.

Ronald Langacker. 1969. *On Pronominalization and the Chain of Command.* In D. Reibel and S. Schane (eds.), Modern Studies in English, Prentice-Hall, Englewood Cliffs, NJ. 160–186.

John Lawler. 2010. Negation and Negative Polarity. *The Cambridge Encyclopedia of the Language Sciences.* Patrick Colm Hogan (ed.) Cambridge University Press. Cambridge, UK.

Andrew MacKinlay, David Martinez and Timothy Baldwin. 2009. Biomedical Event Annotation with CRFs and Precision Grammars. *BioNLP'09: Proceedings of the Workshop on BioNLP*. 77-85.

James McCawley. 1993. *Everything that Linguists have Always Wanted to Know about Logic But Were Ashamed to Ask*. 2nd edition. The University of Chicago Press. Chicago, IL.

David McClosky, Eugene Charniak, and Mark Johnson. 2006. Effective Self-Training for Parsing. *Proceedings of HLT/NAACL 2006.* 152-159.

Roser Morante and Walter Daelemans. 2009. A Metalearning Approach to Processing the Scope of Negation. *CoNLL '09: Proceedings of the 13th Conference on Computational Natural Language Learning.* 21-29.

Pradeep Mutalik, Aniruddha Deshpande, and Prakash M. Nadkarni. 2001. Use of General-purpose Negation Detection to Augment Concept Indexing of Medical Documents: A Quantitative Study using the UMLS. *Journal of the American Medical Informatics Association : JAMIA*. 8(6):598-609.

Jeyakumar Natarajan, Berrar, D., Dubitzky, W., Hack, C., Zhang, Y., DeSesa, C., Van Brocklyn, J. R. and Bremer, E.G. 2006. Text mining of full-text journal articles combined with gene expression analysis reveals a relationship between sphingosine-1-phosphate and invasiveness of a glioblastoma cell line. *BMC Bioinformatics.* 7: 373.

Jon Patrick, Yefeng Wang, and Peter Budd. 2006. Automatic Mapping Clinical Notes to Medical Terminologies. *Proc. Of the 2006 Australian Language Technology Workshop.* 75-82.

Olivia Sanchez. 2007. *Text mining applied to biological texts: beyond the extraction of protein-protein interactions.* PhD Thesis.

Farzaneh Sarafraz, James Eales, Reza Mohammadi, Jonathan Dickerson, David Robertson and Goran Nenadic. 2009. Biomedical Event Detection using Rules, Conditional Random Fields and Parse Tree Distances. *BioNLP'09: Proceedings of the Workshop on BioNLP*.

Herman Tolentino, Michael Matters, Wikke Walop, Barbara Law, Wesley Tong, Fang Liu, Paul Fontelo, Katrin Kohl, and Daniel Payne. 2006. Concept Negation in Free Text Components of Vaccine Safety Reports. *AMIA Annual Symposium proc.*

Yoshimasa Tsuruoka, and Jun'ichi Tsujii. 2005. Bidirectional Inference with the Easiest-First Strategy for Tagging Sequence Data. *Proceedings of HLT/EMNLP 2005*, 467-474.

Contradiction-Focused Qualitative Evaluation of Textual Entailment

Bernardo Magnini
FBK-Irst
Trento, Italy
magnini@fbk.eu

Elena Cabrio
FBK-Irst, University of Trento
Trento, Italy
cabrio@fbk.eu

Abstract

In this paper we investigate the relation between positive and negative pairs in Textual Entailment (TE), in order to highlight the role of contradiction in TE datasets. We base our analysis on the decomposition of Text-Hypothesis pairs into *monothematic pairs*, i.e. pairs where only one linguistic phenomenon at a time is responsible for entailment judgment and we argue that such a deeper inspection of the linguistic phenomena behind textual entailment is necessary in order to highlight the role of contradiction. We support our analysis with a number of empirical experiments, which use current available TE systems.

1 Introduction

Textual Entailment (TE) (Dagan et al., 2009) provides a powerful and general framework for applied semantics. TE has been exploited in a series of evaluation campaigns (RTE - Recognizing Textual Entailment) (Bentivogli et al., 2009), where systems are asked to automatically judge whether the meaning of a portion of text, referred as Text (*T*), entails the meaning of another text, referred as Hypothesis (*H*).

RTE datasets have been mainly built with the purpose of showing the applicability of the TE framework to different semantic applications in Computational Linguistics. Starting from 2005, *[T,H]* pairs were created including samples from summarization, question answering, information extraction, and other applications. This evaluation provides useful cues for researchers and developers aiming at the integration of TE components in larger applications (see, for instance, the use of a TE engine for question answering in the QALL-ME project system[1], the use in relation extraction (Romano et al., 2006), and in reading comprehension systems (Nielsen et al., 2009)).

Although the RTE evaluations showed progresses in TE technologies, we think that there is still large room for improving qualitative analysis of both the RTE datasets and the system results. In particular, we intend to focus this paper on contradiction judgments and on a deep inspection of the linguistic phenomena that determine such judgments. More specifically, we address two distinguishing aspects of TE: (i) the variety of linguistic phenomena that are relevant for contradiction and how their distribution is represented in RTE datasets; (ii) the fact that in TE it is not enough to detect the polarity of a sentence, as in traditional semantic analysis, but rather it is necessary to analyze the dependencies between two sentences (i.e. the *[T,H]* pair) in order to establish whether a contradiction holds between the pair. Under this respect we are interested to investigate both how polarity among Text and Hypothesis affects the entailment/contradiction judgments and how different linguistic phenomena interact with polarity (e.g. whether specific combinations of phenomena are more frequent than others).

As an example, let us consider the pair:

T: Mexico's new president, Felipe Calderon, seems to be doing all the right things in cracking down on Mexico's drug traffickers.[...]

H: Felipe Calderon is the outgoing President of Mexico.

In order to detect the correct contradiction judgment between T and H it is necessary to solve the semantic inference that being the new President of a country is not compatible with being the outgoing President of the same country. This kind of inference requires that (i) the semantic opposition is detected, and that (ii) such opposition is consid-

[1] http://qallme.fbk.eu/

Proceedings of the Workshop on Negation and Speculation in Natural Language Processing, pages 86–94,
Uppsala, July 2010.

		Text snippet (pair 125)	Phenomena	Judg.
T		**Mexico's new president, Felipe Calderon**, seems to be doing all the right things in cracking down on Mexico's drug traffickers. [...]		
H		Felipe Calderon is the outgoing President of Mexico.	lexical:semantic-opposition syntactic:argument-realization syntactic:apposition	C
	H1	Mexico's **outgoing** president, Felipe Calderon, seems to be doing all the right things in cracking down on Mexico's drug traffickers. [...]	lexical:semantic-opposition	C
	H2	**The new president of Mexico**, Felipe Calderon, seems to be doing all the right things in cracking down on Mexico's drug traffickers. [...]	syntactic:argument-realization	E
	H3	**Felipe Calderon** is Mexico's **new president**.	syntactic:apposition	E

Table 1: Application of the decomposition methodology to an original RTE pair

ered relevant for the contradiction judgment in the specific context of the pair.

In order to address the issues above, we propose a methodology based on the decomposition of *[T,H]* pairs into *monothematic pairs*, each representing one single linguistic phenomenon relevant for entailment judgment. Then, the analysis is carried out both on the original *[T,H]* pair and on the monothematic pairs originated from it. In particular, we investigate the correlations on positive and on negative pairs separately, and we show that the strategies adopted by the TE systems to deal with phenomena contributing to the entailment or to the contradiction judgment come to light when analyzed using qualitative criteria. We have experimented the decomposition methodology over a dataset of pairs, which either are marked with a contradiction judgment, or show a polarity phenomenon (either in T or H) which, although present, is not relevant for cotradiction.

The final goal underlying our analysis of contradiction in current RTE datasets is to discover good strategies for systems to manage contradiction and, more generally, entailment judgments. To this aim, in Section 5 we propose a comparison between two systems participating at the last RTE-5 campaign and try to analyze their behaviour according to the decomposition into monothematic pairs.

The paper is structured as follows. Section 2 presents the main aspects related to contradiction within the RTE context. Section 3 explains the procedure for the creation of monothematic pairs starting from RTE pairs. Section 4 describes the experimental setup of our pilot study, as well as the results of the qualitative analysis. Section 5 outlines the preliminary achievements in terms of comparison of systems' strategies in order to manage contradiction. Finally, Section 6 reports on previous work on contradiction and textual entailment.

2 Contradiction and Textual Entailment

In RTE, two kinds of judgment are allowed: two ways (*yes* or *no* entailment) or three way judgment. In the latter, systems are required to decide whether the hypothesis is entailed by the text (*entailment*), contradicts the text (*contradiction*), or is neither entailed by nor contradicts the text (*unknown*). The RTE-4 and RTE-5 datasets are annotated for a 3-way decision: entailment (50% of the pairs), unknown (35%), contradiction (15%). This distribution among the three entailment judgments aims at reflecting the natural distribution of entailment in a corpus, where the percentage of text snippets neither entailing nor contradicting each other is higher than the contradicting ones. Even if this balance seems artificial since in a natural setting the presence of unknown pairs is much higher than the other two judgments (as demonstrated in the Pilot Task proposed in RTE-5 (Bentivogli et al., 2009)), the reason behind the choice of RTE organizers is to maintain a trade-off between the natural distribution of the data in real documents, and the creation of a dataset balanced beween positive and negative examples (as in two way task).

As already pointed out in (Wang, 2009), the similarity between T's and H's in pairs marked as entailment and contradiction is much higher with respect to the similarity between T's and H's in pairs marked as unknown. To support this intuition, (Bentivogli et al., 2009) provides some data on the lexical overlap between T's and H's in the last RTE Challenges. For instance, in RTE-4 the lexical overlap is 68.95% in entailment pairs, 67.97% in contradiction pairs and only 57.36% in

the unknown pairs. Similarly, in RTE-5 the lexical overlap between T's and H's is 77.14% in entailment pairs, 78.93% in contradiction pairs and only 62.28% in the unknown pairs.

For this reason, for contradiction detection it is not sufficient to highlight mismatching information between sentences, but deeper comprehension is required. For applications in information analysis, it can be very important to detect incompatibility and discrepancies in the description of the same event, and the contradiction judgment in the TE task aims at covering this aspect. More specifically, in the RTE task the contradiction judgment is assigned to a T,H pair when the two text fragments are extremely unlikely to be true simultaneously.

According to Marneffe *et al.* (2008), contradictions may arise from a number of different constructions, defined in two primary categories: *i)* those occurring via antonymy, negation, and numeric mismatch, and *ii)* contradictions arising from the use of factive or modal words, structural and subtle lexical contrasts, and world knowledge. Comparing the distribution of contradiction types for RTE-3 and the real contradiction corpus they created collecting contradiction "in the wild" (e.g. from newswire, Wikipedia), they noticed that in the latter there is a much higher rate of negations, numeric and lexical contradictions with respect to RTE dataset, where contradictions of category *(ii)* occur more frequently. Analyzing RTE data of the previous challenges, we noticed that the tendency towards longer and more complex sentences in the datasets in order to reproduce more realistic scenarios, is also reflected in more complex structures determining contradictions. For instance, contradictions arising from overt negation as in (pair 1663, RTE-1 test set):

T: All residential areas in South Africa are segregated by race and no black neighborhoods have been established in Port Nolloth.

H: Black neighborhoods are located in Port Nolloth.

are infrequent in the datasets of more recent RTE challenges. For instance, in RTE-5 test set, only in 4 out of 90 contradiction pairs an overt negation is responsible for the contradiction judgment. In agreement with (Marneffe et al., 2008), we also remarked that most of the contradiction involve numeric mismatch, wrong appositions, entity mismatch and, above all, deeper inferences depending on background and world knowledge, as in (pair 567, RTE-5 test set):

T: "[...] we've done a series of tests on Senator Kennedy to determine the cause of his seizure. He has had no further seizures, remains in good overall condition, and is up and walking around the hospital".

H: Ted Kennedy is dead.

These considerations do not mean that overt negations do not appear in the RTE pairs. On the contrary, they are often present in T,H pairs, but most of the times their presence is irrelevant in the assignment of the correct entailment judgment to the pair. For instance, the scope of the negation can be a phrase or a sentence with additional information with respect to the relevant parts of T and H that allow to correctly judge the pair. This fact could be misleading for systems that do not correcly exploit syntactic information, as the experiments using Linear Distance described in (Cabrio et al., 2008).

3 Decomposing RTE pairs

The qualitative evaluation we propose takes advantage of previous work on monothematic datasets. A *monothematic pair* (Magnini and Cabrio, 2009) is defined as a *[T,H]* pair in which a certain phenomenon relevant to the entailment relation is highlighted and isolated. The main idea is to create such monothematic pairs on the basis of the phenomena which are actually present in the original RTE pairs, so that the actual distribution of the linguistic phenomena involved in the entailment relation emerges.

For the decomposition procedure, we refer to the methodology described in (Bentivogli et al., 2010), consisting of a number of steps carried out manually. The starting point is a *[T,H]* pair taken from one of the RTE datasets, that should be decomposed in a number of monothematic pairs $[T, H_i]_{mono}$, where T is the original Text and H_i are the Hypotheses created for each linguistic phenomenon relevant for judging the entailment relation in *[T,H]*.

In detail, the procedure for the creation of monothematic pairs is composed of the following steps:

1. Individuate the linguistic phenomena which contribute to the entailment in *[T,H]*.

2. For each phenomenon *i*:

(a) Individuate a general entailment rule r_i for the phenomenon *i*, and instantiate the rule using the portion of T which expresses *i* as the left hand side (LHS) of the rule, and information from H on *i* as the right hand side (RHS) of the rule.

(b) Substitute the portion of T that matches the LHS of r_i with the RHS of r_i.

(c) Consider the result of the previous step as H_i, and compose the monothematic pair $[T, H_i]_{mono}$. Mark the pair with phenomenon *i*.

3. Assign an entailment judgment to each monothematic pair.

Relevant linguistic phenomena are grouped using both fine-grained categories and broader categories. Macro categories are defined referring to widely accepted linguistic categories in the literature (e.g. (Garoufi, 2007)) and to the inference types typically addressed in RTE systems: *lexical*, *syntactic*, *lexical-syntactic*, *discourse* and *reasoning*. Each macro category includes fine-grained phenomena (Table 2 reports a list of some of the phenomena detected in RTE-5 dataset).

Table 1 shows an example of the decomposition of a RTE pair (marked as *contradiction*) into monothematic pairs. At step 1 of the methodology both the phenomena that preserve the entailment and the phenomena that break the entailment rules causing a contradiction in the pair are detected, i.e. argument realization, apposition and semantic opposition (column *phenomena* in the table). While the monothematic pairs created basing on the first two phenomena preserve the entailment, the semantic opposition generates a contradiction (column *judgment*).

As an example, let's apply step by step the procedure to the phenomenon of semantic opposition. At step 2a of the methodology the general rule:

Pattern: x $\Leftarrow$ / $\Rightarrow$ y
Constraint: *semantic opposition(y,x)*

is instantiated (*new*$\Leftarrow$ / $\Rightarrow$*outgoing*), and at step 2b the substitution in T is carried out (*Mexico's outgoing president, Felipe Calderon [...]*). At step 2c a negative monothematic pair T, H_1 is composed (column *text snippet* in the table) and marked as *semantic opposition* (macro-category *lexical*), and the pair is judged as *contradiction*.

In (Bentivogli et al., 2010), critical issues concerning the application of such procedure are discussed in detail, and more examples are provided. Furthermore, a pilot resource is created, composed of a first dataset with 60 pairs from RTE-5 test set (30 *positive*, and 30 *negative* randomly extracted examples), and a dataset composed of all the monothematic pairs derived by the first one following the procedure described before. The second dataset is composed of 167 pairs (134 *entailment*, 33 *contradiction* examples, considering 35 different linguistic phenomena).[2]

4 Analysis and discussion

Our analysis has been carried out taking advantage of the pilot resource created by Bentivogli *et al.* (2010). From their first dataset we extracted a sample of 48 pairs ($[T, H]_{sample-contr}$) composed of 30 *contradiction* pairs and 18 *entailment* pairs, the latter containing either in T or in H a directly or an indirectly licensed negation.[3] Furthermore, a dataset of 129 monothematic pairs (96 *entailment* and 33 *contradiction* examples), i.e. $[T, H]_{mono-contr}$, was derived by the pairs in $[T, H]_{sample-contr}$ applying the procedure described in Section 3. The linguistic phenomena isolated in the monothematic pairs (i.e. considered relevant to correctly assign the entailment judgment to our sample) are listed in Table 2.

In RTE datasets only a subpart of the potentially problematic phenomena concerning negation and negative polarity items is represented. At the same time, the specificity of the task lies in the fact that it is not enough to find the correct representation of the linguistic phenomena underlying a sentence meaning, but correct inferences should be derived from the relations that these phenomena contribute to establish between two text fragments. The mere presence of a negation in T is not relevant for the TE task, unless the scope of the negation (a token or a phrase) is present as non-negated in H

[2] Both datasets are freely available at http://hlt.fbk.eu/en/Technology/TE_Specialized_Data

[3] Following (Harabagiu et al., 2006) overt (directly licensed) negations include *i)* overt negative markers such as *not, n't*; *ii)* negative quantifiers as *no*, and expressions such as *no one* and *nothing*; *iii)* strong negative adverbs like *never*. Indirectly licensed negations include: *i)* verbs or phrasal verbs (e.g. *deny*, *fail*, *refuse*, *keep from*); *ii)* prepositions (e.g. *without*, *except*); weak quantifiers (e.g. *few*, *any*, *some*), and *iv)* traditional negative polarity items (e.g. *a red cent* or *anymore*).

phenomena	# pairs $[T,H]_{RTE5-mono-contr}$			
	entailment		contradiction	
	# mono	probab.	# mono	probab.
lex:identity	1	0.25	3	0.75
lex:format	2	1	-	-
lex:acronymy	1	1	-	-
lex:demonymy	1	1	-	-
lex:synonymy	6	1	-	-
lex:semantic-opp.	-	-	3	1
lex:hypernymy	2	1	-	-
TOT lexical	13	0.68	6	0.32
lexsynt:transp-head	2	1	-	-
lexsynt:verb-nom.	6	1	-	-
lexsynt:causative	1	1	-	-
lexsynt:paraphrase	2	1	-	-
TOT lexical-syntactic	11	1	-	-
synt:negation	-	-	1	1
synt:modifier	3	0.75	1	0.25
synt:arg-realization	4	1	-	-
synt:apposition	9	0.6	6	0.4
synt:list	1	1	-	-
synt:coordination	2	1	-	-
synt:actpass-altern.	4	0.67	2	0.33
TOT syntactic	23	0.7	10	0.3
disc:coreference	16	1	-	-
disc:apposition	2	1	-	-
disc:anaphora-zero	3	1	-	-
disc:ellipsis	3	1	-	-
disc:statements	1	1	-	-
TOT discourse	25	1	-	-
reas:apposition	1	0.5	1	0.5
reas:modifier	2	1	-	-
reas:genitive	1	1	-	-
reas:meronymy	1	0.5	1	0.5
reas:quantity	-	-	5	1
reas:spatial	1	1	-	-
reas:gen-inference	18	0.64	10	0.36
TOT reasoning	24	0.59	17	0.41
TOT (all phenomena)	96	0.74	33	0.26

Table 2: Occurrences of linguistic phenomena in TE contradiction pairs

(or viceversa), hence a contradiction is generated. For this reason, 18 pairs of $[T,H]_{sample-contr}$ are judged as *entailment* even if a negation is present, but it is not relevant to correctly assign the entailment judgment to the pair as in (pair 205, RTE-5 test set):

T: A team of European and American astronomers say that a recently discovered extrasolar planet, located ***not*** *far from Earth, contains oceans and rivers of hot solid water. The team discovered the planet, Gliese 436 b [...].*

H: Gliese 436 b was found by scientists from America and Europe.

As showed in Table 2, only in one pair of our sample the presence of a negation is relevant to assign the *contradiction* judgment to the pair. In the pairs we analyzed, contradiction mainly arise from quantity mismatching, semantic opposition (antonymy), mismatching appositions (e.g. *the Swiss Foreign Minister x* contradicts *y is the Swiss Foreign Minister*), and from general inference (e.g. *x became a naturalized citizen of the U.S.* contradicts *x is born in the U.S.*). Due to the small sample we analyzed, some phenomena appear rarely, and their distribution can not be considered as representative of the same phenomenon in a natural setting. In 27 out of 30 contradiction pairs, only one monothematic pair among the ones derived from each example was marked as contradiction, meaning that on average only one linguistic phenomenon is responsible for the contradiction judgment in a TE original pair. Hence the importance of detecting it.

Given the list of the phenomena isolated in $[T,H]_{mono-contr}$ with their frequency both in monothematic positive pairs and monothematic negative pairs, we derived the probability of linguistic phenomena to contribute more to the assignment of a certain judgment than to another (column *probab.* in Table 2). Such probability *P* of a phenomenon *i* to appear in a positive (or in a negative) pair is calculated as follows:

$$P(i|[T,H]_{positive}) = \frac{\#(i|[T,H]_{RTE5-positive-mono})}{\#(i|[T,H]_{RTE5-mono})} \quad (1)$$

For instance, if the phenomenon *semantic opposition* appears in 3 pairs of our sample and all these pairs are marked as *contradiction*, we assign a probability of 1 to a pair containing a semantic opposition to be marked as contradiction. If the phenomenon *apposition* (syntax) appears in 9 monothematic positive pairs and in 6 negative pairs, that phenomenon has a probability of 0.6 to appear in positive examples and 0.4 to appear in negative examples. Due to their nature, some phenomena are strongly related to a certain judgment (e.g. semantic opposition), while other can appear both in positive and in negative pairs. Learning such correlations on larger datasets could be an interesting feature to be exploited by TE systems in the assignment of a certain judgment if the phenomenon i is detected in the pair.

Table 3 reports the cooccurrences of the linguistic phenomena relevant to inference in the pairs marked as *contradiction*. On the first horizontal row all the phenomena that at least in one pair determine contradiction are listed, while in the first column there are all the phenomena cooccurring with them in the pairs. The idea undelying this table is to understand if it is possible to identify recurrent patterns of cooccurrences between phenomena in contradiction pairs. As can be noticed, almost all phenomena occur together with expressions requiring deeper inference

	lex:identity	lex:sem_opposition	synt:negation	synt:modifier	synt:apposition	synt:actpass_altern	reas:meronymy	reas:quantity	reas:gen_inference
lex:identity								1	1
lex:format								1	
lex:acronymy					1				
lex:synonymy	1					1		1	1
lex:hypernymy								1	
lexsynt:vrb-nom	1	1						1	
lexsynt:caus.								1	
synt:modifier									1
synt:arg-realiz.		1							1
synt:apposition		2							3
synt:coord.								1	
synt:actpass	1	1							
disc:coref.	3				1				4
disc:apposition									
disc:anaph-0							1	1	
disc:ellipsis	1	1							2
disc:statements									1
reas:genitive			1						
reas:meronymy								1	
reas:gen-infer.	1			1	3		1	2	1

Table 3: Cooccurrencies of phenomena in contradiction pairs

(*reas:general_inference*), but this is due to the fact that this category is the most frequent one. Beside this, it seems that no specific patterns can be highlighted, but it could be worth to extend this analysis increasing the number of pairs of the sample.

5 Comparing RTE systems' behaviour on contradiction pairs

As introduced before, from a contradiction pair it is possible to extract on average 3 monothematic pairs (Bentivogli et al., 2009), and only one of these monothematic pairs is marked as contradiction. This means that on average only one linguistic phenomenon is responsible for the contradiction judgment in a RTE pair, while the others maintain the entailment relation (i.e. it is possible to correcly apply an entailment rule as exemplified in Section 3). On the contrary, in a pair judged as entailment, all the monothematic pairs derived from it are marked as entailment.

These observations point out the fact that if a TE system is able to correctly isolate and judge the phenomenon that generates the contradiction, the system should be able to assign the correct judgment to the original contradiction pair, despite possible mistakes in handling the other phenomena present in that pair.

In order to understand how it is possible to take advantage of the data analyzed so far to improve a TE system, we run two systems that took part into the last RTE challenge (RTE-5) on $[T, H]_{mono-contr}$.

The first system we used is the EDITS system (Edit Distance Textual Entailment Suite) (Negri et al., 2009)[4], that assumes that the distance between *T* and *H* is a characteristics that separates the positive pairs, for which entailment holds, from the negative pairs, for which entailment does not hold (it is developed according to the two way task). It is based on edit distance algorithms, and computes the *[T,H]* distance as the overall cost of the edit operations (i.e. *insertion*, *deletion* and *substitution*) that are required to transform *T* into *H*. In particular, we applied the model that produced EDITS best run at RTE-5 (acc. on RTE-5 test set: 60.2%). The main features of this run are: Tree Edit Distance algorithm on the parsed trees of *T* and *H*, Wikipedia lexical entailment rules, and PSO optimized operation costs, as described in (Mehdad et al., 2009).

The other system used in our experiments is VENSES[5] (Delmonte et al., 2009), that obtained performances similar to EDITS at RTE-5 (acc. on test set: 61.5%). VENSES applies a linguistically-based approach for semantic inference, composed of two main components: *i)* a grammatically-driven subsystem that validates the well-formedness of the predicate-argument structure and works on the output of a deep parser producing augmented (i.e. fully indexed) head-dependency structures; and *ii)* a subsystem that detects allowed logical and lexical inferences basing on different kind of structural transformations intended to produce a semantically valid meaning correspondence. The system has a pronominal binding module that works at text/hypothesis level separately for lexical personal, possessive and reflexive pronouns, which are substituted by the heads of their antecedents. Also in this case, we applied the same configuration of the system used in RTE evaluation.

Table 4 reports EDITS and VENSES accuracies on the monothematic pairs of $[T, H]_{mono-contr}$.

As said before, the accuracy reported for some very rare phenomena cannot be considered completely reliable. Nevertheless, from these data the main features of the systems can be identified. For instance, EDITS obtains the highest accuracies on the positive monothematic pairs, while it seems it has no peculiar strategies to deal with phenomena

[4] http://edits.fbk.eu/

[5] http://project.cgm.unive.it/venses_en.html

phenomena	EDITS % acc.		VENSES % acc.	
	pos.	neg.	pos.	neg.
lex:identity	100	0	100	33.3
lex:format	100	-	100	-
lex:acronymy	100	-	0	-
lex:demonymy	100	-	100	-
lex:synonymy	80.3	-	80.3	-
lex:semantic-opp.	-	0	-	100
lex:hypernymy	100	-	100	-
TOT lexical	96.7	0	80	66.6
lexsynt:transp-head	100	-	50	-
lexsynt:verb-nom.	83.3	-	16	-
lexsynt:causative	100	-	100	-
lexsynt:paraphrase	100	-	100	-
TOT lexical-syntactic	95.8	-	66.5	-
synt:negation	-	0	-	0
synt:modifier	100	0	33.3	100
synt:arg-realization	100	-	50	-
synt:apposition	100	33.3	55.5	83.3
synt:list	100	-	100	-
synt:coordination	100	-	50	-
synt:actpass-altern.	100	0	25	50
TOT syntactic	100	22.2	52.3	77.7
disc:coreference	95	-	50	-
disc:apposition	100	-	0	-
disc:anaphora-zero	100	-	33.3	-
disc:ellipsis	100	-	33.3	-
disc:statements	100	-	0	-
TOT discourse	99	-	23.3	-
reas:apposition	100	0	100	100
reas:modifier	50	-	100	-
reas:genitive	100	-	100	-
reas:meronymy	100	0	100	0
reas:quantity	-	0	-	80
reas:spatial	100	-	0	-
reas:gen-inference	87.5	50	37.5	90
TOT reasoning	89.5	35.2	72.9	82.3
TOT (all phenomena)	96.2	25	59	81.2

Table 4: RTE systems' accuracy on phenomena

that generally cause contradiction (e.g. *semantic opposition*, *negation*, and *quantity mismatching*). On the contrary, VENSES shows an opposite behaviour, obtaining the best results on the negative cases. Analysing such data it is possible to hypothesize systems' behaviours: for example, on the monothematic dataset EDITS produces a pretty high number of false positives, meaning that for this system if there are no evidences of contradiction, a pair should be marked as entailment (in order to improve such system, strategies to detect contradiction pairs should be thought). On the contrary, VENSES produces a pretty high number of false negatives, meaning that if the system is not able to find evidences of entailment, it assigns the contradiction value to the pairs (for this system, being able to correctly detect all the phenomena contributing to entailment in a pair is fundamental, otherwise it will be marked as contradiction).

6 Related Work

Condoravdi *et al.* (2003) first proposed contradiction detection as an important NLP task, then (Harabagiu et al., 2006) provided the first empirical results for it, focusing on contradiction caused by negation, antonymy, and paraphrases. Voorhees (2008) carries out an analysis of RTE-3 extended task, examining systems' abilities to detect contradiction and providing explanations of their reasoning when making entailment decisions.

Beside defining the categories of construction from which contradiction may arise, Marneffe *et al.* (2008) provide the annotation of the RTE datasets (RTE-1 and RTE-2) for contradiction. Furthermore, they also collect contradiction "in the wild" (e.g. from newswire, Wikipedia) to sample naturally occurring ones.[6]

Ritter *et al.* (2008) extend (Marneffe et al., 2008)'s analysis to a class of contradiction that can only be detected using backgroud knowledge, and describe a case study of contradiction detection based on functional relations. They also automatically generate a corpus of seeming contradiction from the Web text.[7]

Furthermore, some of the systems presented in the previous editions of the RTE challenges attempted specic strategies to focus on the phenomenon of negation. For instance, (Snow et al., 2006) presents a framework for recognizing textual entailment that focuses on the use of syntactic heuristics to recognize false entailment. Among the others, heuristics concerning negation mismatch and antonym match are defined. In (Tatu et al., 2007) the logic representation of sentences with negated concepts was altered to mark as negated the entire scope of the negation. (Ferrandez et al., 2009) propose a system facing the entailment recognition by computing shallow lexical deductions and richer inferences based on semantics, and features relating to negation are extracted. In (Iftene et al., 2009) several rules are extracted and applied to detect contradiction cases.

7 Conclusion

We have proposed a methodology for the qualitative analysis of TE systems focusing on contradiction judgments and on the linguistic phenomena that determine such judgments. The methodology is based on the decomposition of *[T,H]* pairs into *monothematic pairs*, each representing one single linguistic phenomenon relevant for entailment

[6] Their corpora are available at http://www-nlp.stanford.edu/projects/contradiction.

[7] Available at http://www.cs.washington.edu/research/ aucontraire/

judgment.

In particular, the phenomena from which contradiction may arise and their distribution in RTE datasets have been highlighted, and a pilot study comparing the performancies of two RTE systems both on monothematic pairs and on the corresponding original ones has been carried out. We discovered that, although the two systems have similar performances in terms of accuracy on the RTE-5 datasets, they show significant differences in their respective abilities to correctly manage different linguistic phenomena that generally cause contradiction. We hope that the analysis of contradiction in current RTE datasets may bring interesting elements to TE system developers to define good strategies to manage contradiction and, more generally, entailment judgments.

8 Acknowledgements

This work has been partially supported by the LiveMemories project (Active Digital Memories of Collective Life) funded by the Autonomous Province of Trento under the call "Major Projects". We would like to thank Professor Rodolfo Delmonte and Sara Tonelli for running the VENSES system on our datasets.

References

Bentivogli, Luisa, Bernardo Magnini, Ido Dagan, Hoa Trang Dang, and Danilo Giampiccolo. 2009. The Fifth PASCAL RTE Challenge. *Proceedings of the TAC 2009 Workshop on Textual Entailment*. To appear. Gaithersburg, Maryland. 17 November.

Bentivogli, Luisa, Elena Cabrio, Ido Dagan, Danilo Giampiccolo, Medea Lo Leggio, and Bernardo Magnini. 2010. Building Textual Entailment Specialized Data Sets: a Methodology for Isolating Linguistic Phenomena Relevant to Inference. *Proceedings of the 7th LREC conference*. Valletta, Malta. 19-21 May.

Cabrio, Elena, Milen Ognianov Kouylekov and Bernardo Magnini, 2008. Combining Specialized Entailment Engines for RTE-4, *Proceedings of the Text Analysis Conference (TAC 2008)*. Gaithersburg, Maryland, USA, 17-18 November.

Condoravdi, Cleo, Dick Crouch, Valeria de Paiva, Reinhard Stolle, and Daniel Bobrow. 2003. Entailment, Intentionality and Text Understanding *Proceedings of the HLT-NAACL 2003 Workshop on Text Meaning*. Edmonton, Alberta, Canada. 31 May.

Dagan, Ido, Bill Dolan, Bernardo Magnini, and Dan Roth. 2009. Recognizing textual entailment: Rational, evaluation and approaches. *Natural Language Engineering (JNLE)*, Volume 15, Special Issue 04, October 2009, pp i-xvii. Cambridge University Press.

De Marneffe, Marie-Catherine, Anna N. Rafferty and Christopher D. Manning. 2008. Finding Contradictions in Text. *Proceedings of ACL-08: HLT*, pages 10391047. Columbus, Ohio, USA, June.

Delmonte, Rodolfo, Sara Tonelli, Rocco Tripodi. 2009. Semantic Processing for Text Entailment with VENSES. *Proceedings of the TAC 2009 Workshop on TE*. Gaithersburg, Maryland. 17 November.

Garoufi, Konstantina. 2007. Towards a Better Understanding of Applied Textual Entailment. *Master Thesis*. Saarland University. Saarbrücken, Germany.

Ferrández, Óscar, Rafael Muñoz, and Manuel Palomar. 2009. Alicante University at TAC 2009: Experiments in RTE. *Proceedings of the TAC 2009 Workshop on Textual Entailment*. Gaithersburg, Maryland. 17 November.

Harabagiu, Sanda, Andrew Hickl, and Finley Lacatusu. 2006. Negation, Contrast and Contradiction in Text Processing. In Proceedings of AAAI-06. Boston, Massachusetts. July 16-20.

Iftene, Adrian, Mihai-Alex Moruz 2009. UAIC Participation at RTE-5. *Proceedings of the TAC 2009 Workshop on Textual Entailment*. To appear. Gaithersburg, Maryland. 17 November.

Magnini, Bernardo, and Elena Cabrio. 2009. Combining Specialized Entailment Engines. *Proceedings of the LTC '09 conference*. Poznan, Poland. 6-8 November.

Mehdad, Yashar, Matteo Negri, Elena Cabrio, Milen Kouylekov, and Bernardo Magnini. 2009. Using Lexical Resources in a Distance-Based Approach to RTE. *Proceedings of the TAC 2009 Workshop on TE*. Gaithersburg, Maryland. 17 November 2009.

Negri, Matteo, Milen Kouylekov, Bernardo Magnini, Yashar Mehdad, and Elena Cabrio. 2009. Towards Extensible Textual Entailment Engines: the EDITS Package. *AI*IA 2009: Emergent Perspectives in Artificial Intelligence*. Lecture Notes in Computer Science, Springer-Verlag, pp. 314-323. 2009.

Nielsen, Rodney D., Wayne Ward, and James H. Martin. 2009. Recognizing entailment in intelligent tutoring systems. In Ido Dagan, Bill Dolan, Bernardo Magnini and Dan Roth (Eds.) *The Journal of Natural Language Engineering, (JNLE)*. 15, pp 479-501. Copyright Cambridge University Press, Cambridge, United Kingdom.

Ritter, Alan, Doug Downey, Stephen Soderland, and Oren Etzioni. 2008. It's a Contradiction - No, it's not: A Case Study using Functional Relations. *Proceedings of 2008 Conference on Empirical Methods*

in Natural Language Processing. Honolulu, Hawaii. 25-27 October.

Romano, Lorenza, Milen Ognianov Kouylekov, Idan Szpektor, Ido Kalman Dagan, and Alberto Lavelli. 2006. Investigating a Generic Paraphrase-Based Approach for Relation Extraction. *Proceedings of EACL 2006*. Trento, Italy. 3-7 April.

Snow, Rion, Lucy Vanderwende, and Arul Menezes. 2006. Effectively using syntax for recognizing false entailment. *Proceedings of the main conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics.* New York, 4-9 June.

Tatu, Marta, Dan I. Moldovan. 2007. COGEX at RTE 3. *Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing.* Prague, Czech Republic, 28-29 June.

Voorhees, Ellen M. 2008. Contradictions and Justifications: Extentions to the Textual Entailment Task. *Proceedings of ACL-08: HLT*. Columbus, Ohio, USA. 15-20 June.

Wang, Rui, and Yi Zhang. 2009. Recognizing Textual Relatedness with Predicate-Argument Structures. *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*. Singapore, 6-7 August.

Discussion Items

- Is there agreement about what exactly do we understand by processing negation and speculation?
- How necessary is it to process negation and speculation for improving the quality of natural language processing tools/applications?
- What kind of information relative to negation and speculation should be annotated in corpora? Does it suffice to annotate the scope of cues? Is it more useful to annotate events and their factuality?
- Are all cases of negation and speculation equally relevant for the accurate extraction of information?
- At what level should the information be annotated: sentence, document?
- How relevant is discourse level annotation for processing negation and speculation? For example, annotating the structure of a document?
- Are there different annotation needs depending on the domain? For example, should a biomedical text be annotated differently than an opinion text?
- Would it be convenient to build a repository of negation/speculation resources (lexicon, guidelines, etc.)?
- Is it feasible to automatically learn the factuality of an event?

Author Index