HLT-NAACL 2004 Workshop: Biolink 2004,
Linking Biological Literature, Ontologies and Databases, pp. 17-24.
Association for Computational Linguistics.

The Language of Bioscience: Facts, Speculations, and Statements in Between

Marc Light
Library and Information Science
Linguistics Department
University of Iowa
Iowa City, IA 52242
marc-light@uiowa.edu

Xin Ying Qiu
Management Sciences
University of Iowa
Iowa City, IA 52242
xin-qiu@uiowa.edu

Padmini Srinivasan
Library and Information Science
Management Sciences
University of Iowa
Iowa City, IA 52242
padmini-srinivasan@uiowa.edu

Abstract

We explore the use of speculative language in MEDLINE abstracts. Results from a manual annotation experiment suggest that the notion of speculative sentence can be reliably annotated by humans. In addition, an experiment with automated methods also suggest that reliable automated methods might also be developed. Distributional observations are also presented as well as a discussion of possible uses for a system that can recognize speculative language.

1 Introduction

The scientific process involves making hypotheses, gathering evidence, using inductive reasoning to reach a conclusion based on the data, and then making new hypotheses. Scientist are often not completely certain of a conclusion. This lack of definite belief is often reflected in the way scientists discuss their work.

In this paper, we focus on expressions of levels of belief: the expressions of hypotheses, tentative conclusions, hedges, and speculations. "Affect" is used in linguistics as a label for this topic. This is not a well-known topic in the field of text processing of bioscience literature. Thus, we present a large number of examples to elucidate the variety and nature of the phenomena. We then return to a discussion of the goals, importance, and possible uses of this research.

1.1 Examples

The sentences in the following box contain fragments expressing a relatively high level of speculation. The level of belief expressed by an author is often difficult to ascertain from an isolated sentence and often the context of the abstract is needed. All examples in the paper are from abstracts available at the Nation Library of Medicine PubMed webpage (currently http://www.ncbi.nlm.nih.gov/PubMed/). The PubMed identifier is provided following each sentence.

Pdcd4 may thus constitute a useful molecular target for cancer prevention. (1131400)

As the GT box has also previously been shown to play a role in gene regulation of other genes, these newly isolated Sp2 and Sp3 proteins might regulate expression not only of the TCR gene but of other genes as well. (1341900)

On the basis of these complementary results, it has been concluded that curcumin shows very high binding to BSA, probably at the hydrophobic cavities inside the protein. (12870844)

Curcumin down-regulates Ki67, PCNA and mutant p53 mRNAs in breast cancer cells, these properties may underlie chemopreventive action. (14532610)

The next examples contain fragments that are speculative but probably less so than those above. (As we will discuss later, it is difficult to agree on levels of speculation.) The containing sentence does

provide some context but the rest of the abstract if not the full text is often necessary along with enough knowledge of field to understand text.

Removal of the carboxy terminus enables ERP to interact with a variety of ets-binding sites including the E74 site, the IgH enhancer pi site, and the lck promoter ets site, suggesting a carboxy-terminal negative regulatory domain. (7909357)

In addition, we show that a component of the Ras-dependent mitogen-activated protein kinase pathway, nerve growth factor-inducible c-Jun, exerts its effects on receptor gene promoter activity most likely through protein-protein interactions with Sp1. (11262397)

Results suggest that one of the mechanisms of curcumin inhibition of prostate cancer may be via inhibition of Akt. (12682902)

The previous examples contain phrases such as *most likely* and *suggesting*, which in these cases, explicitly mark a level of belief less than 100%. The next examples are not as explicitly marked: *to date* and *such as* can also be used in purely definite statements.

To date, we find that the signaling pathway triggered by each type of insult is distinct. (10556169)

However, the inability of IGF-1, insulin and PMA to stimulate 3beta-HSD type 1 expression by themselves in the absence of IL-4 indicates that the multiple pathways downstream of IRS-1 and IRS-2 must act in cooperation with an IL-4-specific signaling molecule, such as the transcription factor Stat6. (11384880)

These findings highlight the feasibility of modulating HO-1 expression during hypothermic storage to confer tissues a better protection to counteract the damage characteristic of organ transplantation. (12927811)

The words *may* and *might* were both used to express speculation in the examples above but are ambiguous between expressing speculation versus possibility. The examples above are speculative and the sentence below expresses a definite statement about two possibilities.

The level of LFB1 binding activity in adenoid-cystic as well as trabecular tumours shows some variation and may either be lower or higher than in the non-tumorous tissue. (7834800)

The sentence below involves the adjective *putative* in an apositive noun phrase modifier, a different syntactic form that in the previous examples. It also clearly shows that the speculative portion is often confined to only a part of the information provided in a sentence.

We report here the isolation of human zinc finger 2 (HZF2), a putative zinc-finger transcription factor, by motif-directed differential display of mRNA extracted from histamine-stimulated human vein endothelial cells. (11121585)

Of course, definite sentences also come in a variety. The definite sentences below vary in topic and form.

Affinity chromatography and coimmunoprecipitation assays demonstrated that c-Jun and T-Ag physically interact with each other. (12692226)

However, NF-kappaB was increased at 3 h while AP-1 (Jun B and Jun D) and CREB were increased at 15 h. (10755711)

We studied the transcript distribution of c-jun, junB and junD in the rat brain. (1719462)

An inclusive model for all steps in the targeting of proteins to subnuclear sites cannot yet be proposed. (11389536)

We have been talking about speculative fragments and speculative sentences. For the rest of the paper, we define a speculative sentence to be one that contains at least one speculative fragment. A definite sentence contains no speculative fragments. In this study we only considered annotations at the sentence level. However, in future work, we plan to work on sub-sentential annotations.

1.2 Goals of our research on speculative speech and possible uses

Our general goal is to investigate speculative speech in bioscience literature and explore how it might be used in HLT applications for bioscientists. A more specific goal is to investigate the use of speculative speech in MEDLINE abstracts because of their accessibility.

There are a number of reasons supporting the importance of understanding speculative speech:

- it makes up a substantial portion of scientific prose (we estimate that 11% of sentences in MEDLINE abstracts contain speculative fragments),
- many researchers are interested in current trends and directions and speculations are likely to be relevant,
- even if definite statements are of primary importance, knowing that a statement is **not** definite, i.e. speculative, is important.

In the following, we expand upon these points in the contexts of i) information retrieval, ii) information extraction, and iii) knowledge discovery.

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.

In the context of information extraction, a similar situation exists. For example, extracting tables of protein-protein interactions would benefit from knowing which interactions were speculative and which were definite.

In the context of knowledge discovery (KR), speculation might play a number of roles. One possibility would be to use current speculative statements about a topic of interest as a seed for the automated knowledge discovery process. For example, terms could be extracted from speculative fragments and used to guide the initial steps of the knowledge discovery process. A less direct but perhaps even more important use is in building test/train datasets for knowledge discovery systems. For example, let us assume that in a 1985 publication we find a speculation about two topics/concepts A and C being related and later in a 1995 document there is a definite statement declaring that A and C *are* connected via B. This pair of statements can then form the basis of a discovery problem. We may use it to test a KR system's ability to predict B as the connecting aspect between A and C and to do this using data prior to the 1995 publication. The same example could also be used differently: KR systems could be assessed on their ability to make a speculation between A and C using data up to 1985 excluding the particular publication making the speculation. In this way such pairs of temporally ordered speculative-definite statements may be of value in KR research. Differentiating between speculative and definite statements is one part of finding such statement pairs.

2 Related work

We know of no work specifically on speculative speech in the context of text processing of bioscience literature. However, some work on information extraction from bioscience literature has dealt with speculative speech. For example, (Friedman et al., 1994) discusses uncertainty and hedging in radiology reports and their system assigns one of five levels of certainty to extracted findings.

Text processing systems in general have focused "factual" language. However, a growing number of researchers have started work on other aspects of language such as expressing opinions, style of writing, etc. For example a human language technology workshop will be held this Spring entitled "Exploring Attitude and Affect in Text: Theories and Applications." (Qu et al., 2004). Previous work along these lines includes (Wilson and Wiebe, 2003). This research focuses on newswire texts and other texts on the topic of politics and current events.

There has been recent work on classifying sentences from MEDLINE abstracts for the categories such as object, background, conclusions (McKnight and Srinivasan, 2003). In addition, early work,

(Liddy, 1988) built text grammars for empirical research abstracts categorized and assigned structure concerning rhetorical roles of the sentences. However, none of this work addresses the speculative vs. definite distinction we are interested in.

There has also been some work on constructing test sets for knowledge discovery. Several researchers have used the discoveries by Swanson and Smalheiser to test their own algorithms. The two problems most commonly used in replication studies (e.g., (Weeber et al., 2001)) are their discovery of a link between Raynauds disease and fish oils (Swanson, 1986) and their discovery of several links between migraine and magnesium (Swanson, 1988). The most comprehensive replication to date is (Srinivasan, 2004) which employs eight Swanson and Smalheiser discoveries as a test bed.

In the remainder of the paper, we describe a manual annotation experiment we performed, give preliminary results on our attempts to automatically annotate sentences as containing speculative fragments, and conclude with comments on possible future work.

3 Manual annotation experiment

In this experiment, four human annotators manually marked sentences as highly speculative, low speculative, or definite.

Some of the questions we hoped to answer with this experiment were: can we characterize what a speculative sentence is (as demonstrated by good inter-annotator agreement), can a distinction between high and low speculation be made, how much speculative speech is there, where are speculative sentences located in the abstract, is there variation across topics?

The annotators were instructed to follow written annotation guidelines which we provide in appendix of this paper. We wanted to explore how well the annotators agreed on relatively abstract classifications such as "requires extrapolation from actual findings" and thus we refrained from writing instructions such as "if the sentence contains a form of *suggest*, then mark it as speculative" into the guidelines.

We chose three topics to work on and used the following Pubmed queries to gather abstracts:

- "gene regulation" AND "transcription factor" AND 1900:2001[edat]
- (crohn's disease OR crohn disease) AND complications[MeSH Subheading] AND hasabstract[text] AND English[Lang] AND (hominidae[MeSH Terms] OR Human[MeSH Terms])
- turmeric OR curcumin OR curcuma

The first topic is gene regulation and is about molecular biology research on transcription factors, promoter regions, gene expression, etc. The second topic is Crohn's disease which is a chronic relapsing intestinal inflammation and has a number of genes (CARD15) or chromosomal loci associated with it. The third topic is turmeric (aka curcumin), a spice widely used in Asia and highly regarded for its curative and analgesic properties. These include the treatment of burns, stomach ulcers and ailments, and various skin diseases. There has been a surge of interest in curcumin over the last decade.

Each abstract set was prepared for annotation as follows: the order of the abstracts was randomized and the abstracts were broken into sentences using Mxterminator (Reynar and Ratnaparkhi, 1997). The following people performed the annotations: Padmini Srinivasan, who has analyzed crohns and turmeric documents for a separate knowledge discover research task, Xin Ying Qiu, who is completely new to all three topics, Marc Light, who has some experience with gene regulation texts (e.g., (Light et al., 2003)), Vladimir Leontiev, who is a research scientist in an anatomy and cell biology department. It certainly would have been preferable to have four experts on the topics do the annotation but this was not possible.

The following manual annotations were performed:

a. 63 gene regulation abstracts (all sentences) by both Leontiev and Light,

b. 47 gene regulation *additional* abstracts (all sentences) by Light,

c. 100 crohns abstracts (last 2 sentences) by both Srinivasan and Qiu,

d. 400 crohns abstracts *additional* (last 2 sentences) by Qiu,

e. 100 turmeric abstracts (all sentences) by Srinivasan,

f. 400 turmeric *additional* abstracts (last 2 sentences) by Srinivasan.

The 63 double annotated gene regulation abstracts (set a) contained 547 sentences. The additional abstracts (set b) marked by Light[1] contained 344 sentences summing to 891 sentences of gene regulation abstracts. Thus, there is an average of almost 9 sentences per gene regulation abstract. The 100 turmeric abstracts (set e) contained 738 sentences. The other sets contain twice as many sentences as abstracts since only the last two sentences where annotated.

The annotation of each sentence was performed in the context of its abstract. This was true even when only the last two sentences where annotated. The annotation guidelines in the appendix were used by all annotators. In addition, at the start of the experiment general issues were discussed but none of the specific examples in the sets a-f.

We worked with three categories Low Speculative, High Speculative, and Definite. All sentences were annotated with one of these. The general idea behind the low speculative level was that the authors expressed a statement in such a way that it is clear that it follows almost directly from results but not quite. There is a small leap of faith. A high speculative statement would contain a more dramatic leap from the results mentioned in the abstract.

Our inter-annotator agreement results are expressed in the following four tables. The first table contains values for the kappa statistic of agreement (see (Siegel and Castellan, 1988)) for the gene regulation data (set a) and the crohns data (set c). Three values were computed: kappa for three-way agreement (High vs. Low vs. Definite), two-way (Speculative vs. Definite) and two-way (High vs. Low). Due to the lack of any sentences marked High in set c, a kappa value for High vs. low (HvsL) is not possible. Kappa scores between 0.6 and 0.8 are generally considered encouraging but not outstanding.

	HvsLvsD	SvsD	HvsL
geneReg	0.53	0.68	0.03
crohns	0.63	0.63	na

[1]Pun intended.

The following two tables are confusion matrices, the first for gene regulation data (set a) and the second for the crohns data (set c).

	H	L	D
H	5	11	5
L	10	26	19
D	3	12	440

	H	L	D
H	0	0	3
L	0	14	3
D	1	7	170

If we consider one of the annotators as defining truth (gold standard), then we can compute precision and recall numbers for the other annotator on finding speculative sentences. If we choose Leontiev and Srinivasan as defining truth, then Light and Qiu receive the scores below.

	precision	recall
Light	0.68	0.78
Qiu	0.70	0.64

As is evident from the confusion matrices, the amount of data that we redundantly annotated is small and thus the kappa numbers are at best to be taken as trends. However, it does seem that the speculative vs. definite distinction can be made with some reliability. In contrast, the high speculation vs. low speculation distinction cannot.

The gene regulation annotations marked by Light (sets a & b using only Light's annotations) can be used to answer questions about the position of speculative fragments in abstracts. Consider the histogram-like table below. The first row refers to speculative sentences and the second to definite. The columns refer to the last sentence of an abstract, the penultimate, elsewhere, and a row sum. The number in brackets is the raw count. Remember that the number of abstracts in sets a & b together is 100.

	last	2nd last	earlier	total
S	57%(57)	23%(23)	6%(45)	14%(125)
D	43%(43)	77%(75)	94%(648)	86%(766)

It is clear that almost all of the speculations come towards the end of the abstract. In fact the final sentence contains a speculation more often than not. In addition, consider the data where all sentences in an abstract were annotated (sets a & b & e, using Light's annotation of a), there were 1456 definitive sentences (89%) and 173 speculative sentence

(11%). Finally, if we consider the last two sentences of all the data (sets a-f), we have 1712 definitive sentences (82%) and 381 speculative sentences (18.20%).

4 Automatic classifier experiment

We decided to explore the ability of an SVM-based text classifier to select speculative sentences from the abstracts. For this the abstracts were first processed using the SMART retrieval system (Salton, 1971) in order to obtain representation vectors (term-based). Alternative representations were tried involving stemming and term weighting (no weights versus TF*IDF weights). Since results obtained were similar we present only results using stemming and no weights.

The classifier experiments followed a 10-fold cross-validation design. We used SVM_{light} package[2] with all settings at default values. We ran experiments in two modes. First, we considered only the last 2 sentences. For this we pooled all hand tagged sentences from the three topic areas (sets a-f). Second, we explored classification on all sentences in the document (sets a,b,e).

If we assume a default strategy as a simple baseline, where the majority decision is always made, then we get an accuracy of 82% for the classification problem on the last two sentences data set and 89% for the all sentences data set. Another baseline option is to use a set of strings and look for them as substrings in the sentences. The following 14 strings were identified by Light while annotating the gene regulation abstracts (sets a&b): *suggest, potential, likely, may, at least, in part, possibl, potential, further investigation, unlikely, putative, insights, point toward, promise, propose*. The automated system then looks for these substrings in a sentence and if found, the sentence is marked as speculative and as definite if not.

In the table below the scores for the three methods of annotation are listed as rows. We give accuracy on the categorization task and precision and recall numbers for finding speculative sentences. The format is precision/recall(accuracy), all as percentages. The Majority method, annotating every sentence as definite, does not receive precision and recall values. The substring method was run on a subset of the datasets where the gene regulation data (sets a&b) was removed. (It performs extremely well on the gene regulation data due to the fact that it was developed on that data.)

	last2	all
SVM	71/39(85)	84/39(92)
Substr	55/80(87)	55/79(95)
Majority	(82)	(89)

Again the results are preliminary since the amount of data is small and the feature set we explored was limited to words. However, it should be noted that both the substring and the SVM systems performs well suggesting that speculation in abstracts is lexically marked but in a somewhat ambiguous fashion. This conclusion is also supported by the fact that neither system used positional features and yet the precision and recall on the all sentence data set is similar to the last two sentences data set.

5 Conclusion and future work

The work presented here is preliminary but promising. It seems that the notion of speculative sentence can be characterized enabling manual annotation. However, we did not manage to characterize the distinction between high and low speculation. In addition, it seems likely that automated systems will be able to achieve useful accuracy. Finally, abstracts seem to include a fair amount of speculative information.

Future work concerning manual annotation would include revising the guidelines, throwing out the High vs. Low distinction, annotating more data, annotating sub-sentential units, annotating the focus of the speculation (e.g., a gene), and annotating full text articles. We are also ignorant of work in linguistics that almost certainly exists and may be informative. We have started this process by considering (Hyland, 1998) and (Harris et al., 1989).

Future work concerning automatic annotation includes expanding the substring system with more substrings and perhaps more complicated regular expressions, expanding the feature set of the SVM, trying out other classification methods such as decision trees.

[2] http://wwwai.cs.unidortmund.de/SOFTWARE/SVM_LIGHT/svm_light.html.en

Finally, we plan on building some of the applications mentioned: a speculation search engine, transcription factor interaction tables with a speculation/definite column, and knowledge discovery test sets.

Acknowledgments

We would like to thank Vladimir Leontiev for his time and effort annotating gene regulation abstracts. In addition, we would like to thank David Eichmann for his assistance with our database queries. We would also like to thank Lynette Hirschman for assistance with the title of this paper. Finally, we would like to thank the anonymous workshop reviewers for their comments.

References

C. Friedman, P. Alderson, J. Austin, J.J. Cimino, and S.B. Johnson. 1994. A general natural-language text processor for clinical radiology. *Journal of the American Medical Informatics Association*, 1(2):161–174.

Z. Harris, M. Gottfried, T. Ryckman, P. Mattick, A. Daladier, T.N. Harris, and S. Harris. 1989. *The Form of Information in Science : analysis of an immunology sublanguage*. Kluwer Academic Publishers.

K. Hyland. 1998. *Hedging in Scientific Research Articles*. John Benjamins B.V.

E. D. Liddy. 1988. *The Discourse-Level Structure of Natural Language Texts: An Exploratory Study of Empirical Abstracts*. Ph.D. thesis, Syracuse University.

M. Light, R. Arens, V. Leontiev, M. Patterson, X. Y. Qiu, and H. Wang. 2003. Extracting transcription factor interactions from medline abstracts. In *Posters from the 11th International Conference on Intelligent Systems in Molecular Biology*. ISCB.

L. McKnight and P. Srinivasan. 2003. Categorization of sentence types in medical abstracts. In *Proceedings of the 2003 AMIA conference*.

Yan Qu, J. Shanahan, and J. M. Wiebe, editors. 2004. *Proceedings of the AAAI Spring Symposium on Exploring Attitude and Affect in Text: Theories and Applications*. AAAI. (to appear).

J. Reynar and A. Ratnaparkhi. 1997. A maximum entropy approach to identifying sentence boundaries. In *Proceedings of the Fifth Conference on Applied Natural Language Processing*, pages 16–19. ACL.

G. Salton, editor. 1971. *The SMART Retrieval System: Experiments in Automatic Document Processing*. Prentice Hall, Englewood Cliffs, NJ.

S. Siegel and N.J. Castellan. 1988. *Nonparametric Statistics for the Behavioral Sciences*. McGraw-Hill.

P. Srinivasan. 2004. Text mining: Generating hypotheses from medline. *Journal of the American Society for Information Science and Technology*. (to appear).

D.R. Swanson. 1986. Fish oil, raynaud's syndrome, and undiscovered public knowledge. *Perspectives in Biology and Medicine*, 30:7–18.

D.R. Swanson. 1988. Migraine and magnesium: Eleven neglected connections. *Perspectives in Biology and Medicine*, 31:526–557.

M. Weeber, H. Klein, L. Berg, and R. Vos. 2001. Concepts in literature-based discovery: Simulating swanson's raynaud-fish oil and migraine-magnesium discoveries. *Journal of the American Society for Information Science*, 52(7):548–557.

T. Wilson and J. Wiebe. 2003. Annotating opinions in the world press. In *Proceedings of the 4th SIGdial Workshop on Discourse and Dialogue (SIGdial-03)*.

Appendix: Annotation Guidelines

Some target uses for speculative sentence classification:

- a speculation search site that enables scientists and health workers to find speculative statements about a topic of interest,
- a set of starting points for knowledge discovery systems,
- a test set for knowledge discovery systems.

The purpose of the guidelines below is to instruct annotators on which sentences should be marked as speculative.

There are three possible annotations for a sentence: Low Speculative, High Speculative, and Definite. All sentences should annotated with one of these.

A sentence may be long and contain many subparts:

- if any part of it is High Speculative (HS), it should be marked as HS,
- if it is not HS but a part of it is Low Speculative (LS), it should be marked as LS,

- otherwise it should be marked as Definite.

It should also be mentioned that the intent of the author is what is relevant. The annotator should try to decide if the author meant the sentence as speculative or definite. E.g., an annotator should not mark a sentence as speculative, if the author intended the statement to be definitive.

Below are the definitions for the categories.

- Low Speculative (LS): A sentence fragment is LS if the author indicates that it receives direct support from the work presented but there are other possible explanations for the results (as there always are in science). However, the proposition (expressed in the sentence fragment) is a plausible if not likely explanation.

- High Speculative (HS): A sentence fragment is HS if the author indicates that it does not follow from the work presented but could be extrapolated from it. In other words the work provides indirect support for the proposition.

- Definite: A sentence fragment is definite if it is not LS or HS. Observations are generally Definite as are statements about methods, previous work, etc.

Below are tests that may be helpful for annotating particular sentences.

- If the sentence fragment implicitly suggests future experimentation, then it is likely to be HS.

- Paraphrased the sentence fragment using “we conclude”, “we observe”, or “we know”. If a contradiction or cognitive dissonance occurs then perhaps the fragment is speculative. The contradiction will be analogous to that in “we definitely believe that maybe there is a chance”.

Below are a number of additional considerations.

- Our characterization of speculative speech is meant to be broad enough to include statements that are not explicitly marked as speculations but are speculations made by the authors nonetheless. For example, we would consider a proposal that some statement is true to be a speculative sentence.

- Mentions of speculations made in previous work should be considered speculations, e.g., “It was recently proposed that ...”.