A Review of Negation in Clinical Texts: DIT Technical Report: SOC-AIG-001-08*

John D. Kelleher and Brian Mac Namee
Artificial Intelligence Group
School of Computing
Dublin Institute of Technology

1 Introduction

Negation is commonly seen in clinical documents [Chapman *et al.*, 2001a]

"In clinical reports the presence of a term does not necessarily indicate the presence of the clinical condition represented by that term. In fact, many of the most frequently described findings and diseases in discharge summaries, radiology reports, history and physical exams, and other transcribed reports are denied in the patient" [Chapman *et al.*, 2001b, page. 301].

1.1 Negation Background

[Huang and Lowe, 2007] decompose a negation into three components: a negation signal, a negated phrase containing one or more concept(s), and optionally some supporting feature (patter), which helps us locate the negated phrase. The illustrate this decomposition using the following example: *There is* **no** *evidence of* cervical lymph node enlargement. In this example, **no** is the negation signal used to denote the following concept is negated; cervical lymph node enlargement is the negated phrase; while *evidence of* is the supporting feature.

Within Huang and Lowe's decomposition, I think that Chapman *et al.*'s term **pertinent negative** refers to the negated phrase containing one or more negated concepts. Chapman *et al.* 2001b note that identifying a pertinent negation involves (see [Chapman *et al.*, 2001b] Section 1.1):

1. identifying a proposition ascribing a clinical condition to a person

*Available online at: http://www.comp.dit.ie/aigroup/

2. determining whether the proposition is denied or negated in the text

1.2 Why negation is difficult

1.2.1 Scope

One of the major problems with handling negation is identifying the scope of a negation signal. [Chapman *et al.*, 2001b] use a simple horizon of 5 words from the negation signal to a UMLS term to define whether or not that term is within the scope of the negation signal. [Huang and Lowe, 2007] use a more sophisticated approach where regular expressions (associated negation signal with patterns) are used to classify a negation, into 1 of 11 classes that they have defined, and then grammatical rules, that have been hand-crafted and associated with each class of negation they have constructed, iterate over a parse tree are used to define the scope of the negation.

[Huang and Lowe, 2007, page. 307] notes: "There are also conjunctions such as *but*, prepositions such as *besides* and adverbs such as *other than* that reduce the scope of a negation within

See [Chapman *et al.*, 2001b, page 307] and [Mutalik *et al.*, 2001] for discussion on determining the scope of **not**.

[Chapman *et al.*, 2001b, page 307] "it is common for a physician to negate several findings or diseases in a comma-separated list."

1.2.2 Types

Another difficulty with negation is that it is a very complex linguistic phenomenon with many different types. Huang and Lowe's 2007 direct the reader to [Mutalik *et al.*, 2001] for discussion (I haven't read [Mutalik *et al.*, 2001] yet). However, in their own Method section highlights some of them:

complete versus partial negations ("probably not" would be a partial negation signal). [Chapman *et al.*, 2001a] note some nouns, such as "change" are not real negation when appearing as the head of a negated noun phrase (reported in [Huang and Lowe, 2007, page. 307]). See also discussion in [Chapman *et al.*, 2001b, page. 303] regarding their "pseudo-negation" category of negation signals.

negation within a word (as in the case of negative prefix or suffix) are often semantically ambiguous. [Huang and Lowe, 2007] note that "the best way to represent these words may depend on the controlled terminology used for concept encoding"; they also note that many UMLS concepts represent

antonymous forms of other concepts - e.g., words beginning with "anti-", "an-", "un-" and "non-" - which are not necessarily negations.

double negation "We **cannot exclude** malignancy" See [Huang and Lowe, 2007] page 307 column 2 and Table 1.

1.2.3 Multiple occurrences of terms in a sentence

How do you handle multiple occurrences of a medical term in a sentence some of which are negated? See [Chapman *et al.*, 2001b] for discussion Section 4.1.

1.2.4 It can extend across sentences

Negation can extend across sentence boundaries. See [Chapman *et al.*, 2001b, page 308] for discussion (including affect of temporal relations on multiple occurrences).

2 Possible Pre-processing Stages

2.1 Mapping free text NP to a controlled terminology

Although [Huang and Lowe, 2007] do not use such a mapping they note that several studies on negation map concepts to a controlled terminology such as UMLS or SNOMED CT: see their background section for a good overview. They note that the mapping can introduce errors.

[Chapman *et al.*, 2001b] use this type of mapping as a pre-processing step. They illustrate this process using an example that involves rewriting the sentence *The patient denied experiencing chest pain on exertion* as *The patient denied experiencing <S1459038> on exertion.* They note that one of the limitations of their system is that they use string matching to identify relevant UMLS phrase and they suggest using more sophisticated methods for indexing documents with UMLS phrase, directing the reader to[Nadkarni *et al.*, 2001] (it may also be worth looking at [Mutalik *et al.*, 2001])

2.2 Co-reference resolution

[Chapman *et al.*, 2001b] note that a limitation of their system is their assumption that a sentence-level analysis is sufficient for identifying pertinent negatives. Although they argue that this is generally a reasonable assumption they note that systems that do ignore co-reference relationships "will probably miss some pertinent negatives because the UMLS term is referred to in another sentence by a pronoun such as *it* or a generic description of the term such as *the finding*".

3 Systems

3.1 NegFinder

[Mutalik *et al.*, 2001]

Uses a lexical scanner with regular expressions and a parser that uses a restricted context-free grammar to identify pertinent negatives in discharge summaries and surgical notes. The system first identifies propositions or concepts and then determines whether the concepts are negated.

Uses a One-token Look-Ahead Left-to-right Rigtmost-derivation (LALR(1)) parser to detect negations in surgical notes and discharge summaries without extracting syntactical structures of sentences and phrases as in full NLP parsing.

Performance:

Sensitivity 95.7%

Specificity 91.8%

3.2 NegEx

[Chapman *et al.*, 2001b]

1. Process one sentence at a time.
2. Remove all punctuation (retain stop words - some commonly used stop words - e.g. *of* are important parts of the expressions we are looking for)
3. Identify UMLS terms in the text and replace with unique string identifiers from the UMLS (using string matching - matching the longest possible string among eligible matches in the UMLS)
4. Use regular expressions to identify negation signals in the text and negate all UMLS terms within a window of 5 words of the negation signal. Note: NegEx defined 3 types of regular expressions: group (1) pseudo-negation; group (2) <negation phrase> * <UMLS term>; group (3) <UMLS term> * <negation phrase> (* indicates 5 tokens - words or UMLS terms 0 may fall between the negation and the UMLS term). Regular expressions were matched to the longest possible subset of the sentence.

A current and updated list of regular expressions and negated phrases used by NegEx is available at:

`http://omega.cbmi.upmc.edu/~chapman/NegEx.html`

NegEx treated multiple occurrences of a term in a sentence as a single occurrence. If a term s is negated at least once then all occurrences of it in that sentence are negated.

Using a human annotated corpus of 500 sentences containing the regular expressions defined in the system the systems performance was:

Sensitivity 82.41%

Specificity 82.5%

PPV 84.49%

NPV 80.21%

3.3 Elkin et al. 2005

[Elkin *et al.*, 2005]

Uses a negation ontology containing operators and their associated rules. Operators were two sets of terms with one set starting negations and another set stopping the propagation of negations.

1. Break each sentence into text and operators
2. Text mapped to SNOMED CT concepts
3. Concepts are then assigned one of three possible assertions according to the negation ontology,.

Performance: Not including negated concepts that could not be mapped to SNOMED CT concepts (205 of 2,028 concepts identified and negated by the human reviewer):

Sensitivity 97.2%

Specificity 98.8%

PPV 91.2%

3.4 ChartIndex

[Huang and Lowe, 2007]

Negations are classified based upon the syntactical categories of negation signals, and negation patterns, using regular expression matching. Negated terms are then located in parse trees using corresponding negation grammar rules.

Performance:

Sensitivity 92.6%

Specificity 99.87%

PPV 98.6%

4 Notes

- Most phrases indicating negation are stop words in information retrieval systems and are not even used for indexing [Chapman *et al.*, 2001a]
- Negation phrases appear to comply qualitatively with Zipf's law. [Chapman *et al.*, 2001b] results indicate that there are a few very common negation phrases (*no*, *without*, *no evidence of*, more medium-frequency negation phrases, and a potentially huge number of low-frequency phrases.
- Clinicians often negate long strings of diseases
- "MEDLINE indexing uses sophisticated syntactic and semantic processing techniques, but does not incorporate explicit distinctions between positive and negative terms" [Mutalik *et al.*, 2001, page. 302]

5 Acronyms and Terminology

5.1 Acronyms

MLP medical language processing

NLM The National Library of Medicine

UMLS Unified Medical Language System: "provides comprehensive coverage of biomedical concepts" [Huang and Lowe, 2007]

SNOMED CT Medical ontology similar in nature to UMLS (see [Huang and Lowe, 2007] background section for discussion)

ICD(10) the International Statistical Classification of Disease and Related Health Problems, 10^{th} revision [Chapman *et al.*, 2001b]

PPV Positive predictive value

NPV Negative predictive value

5.2 Terminology

Pertinent negative "We use the term *pertinent negative* to refer to findings and diseases explicitly or implicitly described as absent in patient" [Chapman *et al.*, 2001b].

Sensitivity $\frac{number-of-terms-correctly-negated}{number-of-terms-negated-by-rater}$ [Chapman *et al.*, 2001b]

Specificity $\frac{number-of-terms-correctly-not-negated}{number-of-terms-not-negated-by-rater}$ [Chapman *et al.*, 2001b]

PPV $\frac{number-of-terms-correctly-negated}{number-of-terms-negated}$ [Chapman *et al.*, 2001b]

NPV $\frac{number-of-terms-correctly-not-negated}{number-of-terms-not-negated}$ [Chapman *et al.*, 2001b]

6 Negation: Keywords, Structures

- A current and updated list of regular expressions and negated phrases used by NegEx is available at:

  `http://omega.cbmi.upmc.edu/~chapman/NegEx.html`

- Terms that [Chapman *et al.*, 2001b, page. 304] suspected might sometimes be used to signal a pertinent negative: "minimal sign of", "nonfocal", "nonspecific", "unremarkable", "failed", "negative", "never", "nor", "unable". [Chapman *et al.*, 2001b] note that approximately 15% of their discharge reports contained one of the phrases described in this list.

- "versus" is sometimes used to indicate ambiguity "... pnemonica versus bronchitas for her cough" (see [Chapman *et al.*, 2001b, page. 307]

- "not": determining the scope of not is complex: 'This is not an infection" v. "This is not the source of the infection" — "We did not treat the infection" v. "We did not detect an infection" – although these sentences have a similar syntactic structure the finding "infection" is present in the patient in the former and absent in the latter sentence. (see [Chapman *et al.*, 2001b, page. 307]. See also [Mutalik *et al.*, 2001] for discussion on not.

- NegEx - missed negations may be caused by: passive syntactic structures: "X was ruled out", extensive modifiers between the negation and the UNLS term.

- NegEx - false positives cause by failure to decrease the scope of a negation, and by failure to distinguish current visits from the patient's past history "no history of previous cva"
- Most common negation phrases: "no", "without", "no evidence of" [Chapman *et al.*, 2001b, page. 308]

References

[Chapman *et al.*, 2001a] W. W. Chapman, W. Bridewell, P. Hanbury, G. F. Cooper, and B. G. Buchanan. Evaluation of negation phrases in narrative clinical reports. *Proc AMIA Symp*, 105:9, 2001.

[Chapman *et al.*, 2001b] W. W. Chapman, W. Bridewell, P. Hanbury, G. F. Cooper, and B. G. Buchanan. A simple algorithm for identifying negated findings and diseases in discharge summaries. *Journal of Biomedical Informatics*, 34:301–310, 2001.

[Elkin *et al.*, 2005] P. L. Elkin, S. H. Brown, B. A. Bauer, C. S. Husser, W. Carruth, L. R. Bergstrom, and D. L. Wahner-Roedler. A controlled trial of automated classification of negation from clinical notes. *feedback*, 2005.

[Huang and Lowe, 2007] Y Huang and HJ Lowe. A novel hybrid approach to automated negation detection in clinical radiology reports. *Journal of the American Medical Informatics Association*, 14:304–311, 2007.

[Mutalik *et al.*, 2001] P.G. Mutalik, A. Deshpande, and P. Nadkarni. 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*, 8:589–609, 2001.

[Nadkarni *et al.*, 2001] P Nadkarni, R Chen, and C Brandt. Umls concept indexing for production databases: a feasibility study. *Journal of the American Medical Informatics Association*, 8:80–91, 2001.