VetCompass: Clinical Natural Language Processing for Animal Health

Timothy Baldwin
Talk Outline

1. The Eight “P”s of Clinical NLP
2. VetCompass
3. The Appeals/Challenges of Clinical NLP
4. Cross-comparison of Social Media and Clinical Text Sources
   - Corpora
   - Intra-Corpus Analysis
   - Inter-Corpus Analysis
   - Summary
5. Concluding Remarks
The Eight “P”s of Clinical NLP

- We all know that NLP works best when we have big, public-domain, clean text corpora, but these have proven elusive for clinical NLP
The Eight “P”s of Clinical NLP

We all know that NLP works best when we have big, public-domain, clean text corpora, but these have proven elusive for clinical NLP ... but why?
The Eight “P”s of Clinical NLP

We all know that NLP works best when we have big, public-domain, clean text corpora, but these have proven elusive for clinical NLP ... but why?

Taking inspiration from the five “V”s that make “big data” challenging [McAfee and Brynjolfsson, 2012]:

- volume
- velocity
- variety
- veracity
- value

... for clinical NLP, we have the Eight “P”s ...
“P” #1: Privacy

- There are all sorts of privacy laws associated with clinical data/EHRs in different constituencies around the world, restricting sharing of the data and governing levels of anonymisation, with serious punitive and reputational consequences for breaches.

- The upshot of this is that clinical organisations tend to be highly risk averse, and unable/unwilling to make data available to researchers, such that even if data is made available to NLP researchers, it is done in a highly restricted manner.
“P” #2: Paper

- It’s a sad fact that, in many countries (incl. Australia!), a lot of language-based clinical data is still handwritten in paper form only, or if it is digitised, it is in audio form or scanned without OCR.

- Here, the issue is not data access/management per se, but data format/accessibility.

Source(s): https://sandythepa.files.wordpress.com/2014/07/doc2.jpg
“P” #3: Parochialism

- Most countries are still a long way from having common standards for electronic health records ("EHRs"), or at least in the uptake of those standards.
- Even if common standards exist, there is a lot of siloing of clinical data between health organisations, due to:
  - legacy systems/standards
  - local legislative requirements
  - commercial pressures (infrastructure investment, commercial sensitivity, ...)
- That is, legislative and institutional data “parochialism” makes it hard to get access to large-scale datasets.
“P” #4: Paucity

Even where NLP researchers have been able to access/make available clinical text data, it has inevitably been small-scale, available only to a small group of researchers, in large part because of all of the above reasons.
... but let’s say we’ve got access to clinical text, what are the challenges then?
“P” #5: Perplexity

- Clinical text serves many purposes, prominent among which is to serve as a “memory-picture” for clinicians [Nygren and Henriksson, 1992]

- In this context, context is as important as content, and there is little incentive for the clinician to make the clinical note accessible to others, inevitably resulting in “perplexing” notes which are hard to interpret for anyone else

- Note that this varies greatly depending on the author; for nursing staff, e.g., clinical notes serve an important function in handover between shifts, where this is an incentive to be intelligible to others [Poissant et al., 2005]
“P” #6: Perspective

Clinical text is generally very goal-oriented — to improve the delivery of healthcare to the patient — and the content is very much biased towards this information need, meaning:

- there can be a lot of subjectivity in what is included in clinical text (observational bias)
- there is a strong bias towards the speciality of the clinician (domain bias)
- there is a strong bias based on who the author feels is likely to read the clinical text (audience bias)

‘wife’ does the F+W so not known
“P” #7: Parsimony

- Medical practitioners are famously time-poor, which has immediate implications for the verbosity (or lack thereof) of clinical text, the care with which clinical notes are generated, the use of abbreviations, ...

- feel everything else OK at mo
“P” #8: Pragmatics

- Clinical notes often form a time-series of observations, hypotheses, test results, follow-up consultation, etc.
- Without discourse structure/pragmatic processing, they can be very difficult to make sense of

result negative
“P” Recap

Recapping the eight “P”s of clinical NLP complexity:

▶ “P”s of data access complexity:
  - privacy
  - paper
  - parochialism
  - paucity

▶ “P”s of data processing complexity:
  - perplexity
  - perspective
  - parsimony
  - pragmatics
Talk Outline

1. The Eight “P”s of Clinical NLP
2. VetCompass
3. The Appeals/Challenges of Clinical NLP
4. Cross-comparison of Social Media and Clinical Text Sources
   - Corpora
   - Intra-Corpus Analysis
   - Inter-Corpus Analysis
   - Summary
5. Concluding Remarks
What is VetCompass?

- Centralised repository of clinical records from veterinary practices
- Started up in UK, where it has collected >40M treatment records for companion animals from ≈500 practices; recently started up in Australia
- Potential applications:
  - investigation of best practice in animal care
  - tracking of animal health as relevant to zoonoses
  - better veterinary teaching practice
The Eight “P”s and VetCompass

How does VetCompass fit in re the eight “P”s?

The four data access complexity “P”s:

▶ **privacy**: still a concern (for different reasons), but legal barriers largely removed
▶ **paper**: almost exclusively natively digital data
▶ **parochialism**: less siloing/better standards compliance/greater appetite for (controlled) data sharing
▶ **paucity**: nope!

The four data processing complexity “P”s:

▶ **perplexity, perspective, parsimony** and **pragmatics** are still very much in play
Why is VetCompass of Interest to NLP?

- Clinical data at scale, with real-world problems requiring NLP expertise
- Similar in style/nature to human clinical data, opening up possibilities for development of models over VetCompass to transfer across to human clinical notes
- Possibly greater treatment skew in the data (desexing, worm/flea/tick treatments, nail clipping, ...) but a very long tail
Talk Outline

1. The Eight “P”s of Clinical NLP
2. VetCompass
3. The Appeals/Challenges of Clinical NLP
4. Cross-comparison of Social Media and Clinical Text Sources
   - Corpora
   - Intra-Corpus Analysis
   - Inter-Corpus Analysis
   - Summary
5. Concluding Remarks
Standard Assumptions Made in NLP Research

- Edited text
- Static data
- Long(ish) documents; plenty of context
- All context is language context
- Well-defined domain/genre
- Sentence tokenisation
- Grammaticality
NLP Challenges in Clinical NLP

- Edited text
NLP Challenges in Clinical NLP

- Unedited text
NLP Challenges in Clinical NLP

- Unedited text
- Static data
NLP Challenges in Clinical NLP

- Unedited text
- Streamed data
NLP Challenges in Clinical NLP

- Unedited text
- Streamed data
- Long(ish) documents; plenty of context
NLP Challenges in Clinical NLP

- Unedited text
- Streamed data
- Short documents; v. little linguistic context
NLP Challenges in Clinical NLP

- Unedited text
- Streamed data
- Short documents; v. little linguistic context
- All context is language context
NLP Challenges in Clinical NLP

• Unedited text
• Streamed data
• Short documents; v. little linguistic context
• Little language, potentially lots of other context
NLP Challenges in Clinical NLP

- Unedited text
- Streamed data
- Short documents; v. little linguistic context
- Little language, potentially lots of other context
- Well-defined domain/genre
NLP Challenges in Clinical NLP

- Unedited text
- Streamed data
- Short documents; v. little linguistic context
- Little language, potentially lots of other context
- Well-defined, highly-specialised domain
NLP Challenges in Clinical NLP

- Unedited text
- Streamed data
- Short documents; v. little linguistic context
- Little language, potentially lots of other context
- Well-defined, highly-specialised domain
- Sentence tokenisation
NLP Challenges in Clinical NLP

- Unedited text
- Streamed data
- Short documents; v. little linguistic context
- Little language, potentially lots of other context
- Well-defined, highly-specialised domain
- What’s a sentence?
NLP Challenges in Clinical NLP

- Unedited text
- Streamed data
- Short documents; v. little linguistic context
- Little language, potentially lots of other context
- Well-defined, highly-specialised domain
- What’s a sentence?
- Grammaticality
NLP Challenges in Clinical NLP

- Unedited text
- Streamed data
- Short documents; v. little linguistic context
- Little language, potentially lots of other context
- Well-defined, highly-specialised domain
- What’s a sentence?
- Yer what?
NLP Challenges in Clinical NLP

- Haven’t I heard you say that in another context?
NLP Challenges in Clinical NLP

- Haven’t I heard you say that in another context?
  ... guilty as charged ... I have said largely the same things about social media elsewhere
NLP Challenges in Clinical NLP

- Haven’t I heard you say that in another context?
  ... guilty as charged ... I have said largely the same things about social media elsewhere
- So clinical NLP is just social media NLP over a different data source?
NLP Challenges in Clinical NLP

- Haven’t I heard you say that in another context?
  ... guilty as charged ... I have said largely the same things about social media elsewhere

- So clinical NLP is just social media NLP over a different data source?
  ... there are notable differences (e.g. language mix, domain specificity), and very different causes for the data looking like it does, but there are certainly clear similarities
NLP Challenges in Clinical NLP

- Haven’t I heard you say that in another context?
  ... guilty as charged ... I have said largely the same things about social media elsewhere
- So clinical NLP is just social media NLP over a different data source?
  ... there are notable differences (e.g. language mix, domain specificity), and very different causes for the data looking like it does, but there are certainly clear similarities
- So what does the data look like, and how does it contrast with social media text?
Haven’t I heard you say that in another context?
   ... guilty as charged ... I have said largely the same things about social media elsewhere
So clinical NLP is just social media NLP over a different data source?
   ... there are notable differences (e.g. language mix, domain specificity), and very different causes for the data looking like it does, but there are certainly clear similarities
So what does the data look like, and how does it contrast with social media text?
   ... glad you asked!
Talk Outline

1. The Eight “P”s of Clinical NLP
2. VetCompass
3. The Appeals/Challenges of Clinical NLP
4. Cross-comparison of Social Media and Clinical Text Sources
   - Corpora
   - Intra-Corpus Analysis
   - Inter-Corpus Analysis
   - Summary
5. Concluding Remarks
Background

- In earlier work [Baldwin et al., 2013], I carried out a cross-comparison of “noise” in different social media sources, focusing on:
In earlier work [Baldwin et al., 2013], I carried out a cross-comparison of “noise” in different social media sources, focusing on:

... how “noisy” are different social media sources, and in what ways?
Background

- In earlier work [Baldwin et al., 2013], I carried out a cross-comparison of “noise” in different social media sources, focusing on:
  - ... how “noisy” are different social media sources, and in what ways?
  - ... are different social media sources differently “noisy” or are they much of a muchness?
Background

In earlier work [Baldwin et al., 2013], I carried out a cross-comparison of “noise” in different social media sources, focusing on:

... how “noisy” are different social media sources, and in what ways?
... are different social media sources differently “noisy” or are they much of a muchness?
... ultimately, are the differences between social media sources all that great?
**Background**

- In earlier work [Baldwin et al., 2013], I carried out a cross-comparison of “noise” in different social media sources, focusing on:
  - how “noisy” are different social media sources, and in what ways?
  - are different social media sources differently “noisy” or are they much of a muchness?
  - ultimately, are the differences between social media sources all that great?
- Let’s add clinical data to the mix and see what we find ...
Outline of Approach

1. Assemble corpora across a spectrum of social media and clinical data sources + BNC
Outline of Approach

1. Assemble corpora across a spectrum of social media and clinical data sources + BNC
2. Apply a range of analyses to each individual social media source
Outline of Approach

1. Assemble corpora across a spectrum of social media and clinical data sources + BNC
2. Apply a range of analyses to each individual social media source
3. Perform comparative analysis between different corpus pairings
Contents

1. The Eight “P”s of Clinical NLP
2. VetCompass
3. The Appeals/Challenges of Clinical NLP
4. Cross-comparison of Social Media and Clinical Text Sources
   - Corpora
     - Intra-Corpus Analysis
     - Inter-Corpus Analysis
     - Summary
5. Concluding Remarks
Social Media Corpora

- Social media sources targeted in this research:
  1. **Twitter**: micro-blog posts from Twitter
  2. **Comments**: comments from YouTube
  3. **Blogs**: blog posts from Spinn3r dataset
  4. **Forums**: forum posts from popular forums

- As a balanced, non-social media counterpoint corpus:
  6. **BNC**: written portion of British National Corpus
Twitter

- 1M posts from garden hose feed of Twitter, in the form of two sub-corpora collected at different times:
  
<table>
<thead>
<tr>
<th></th>
<th>Twitter-1</th>
<th></th>
<th>Twitter-2</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>22 Sep, 2011</td>
<td>B</td>
<td>22 Feb, 2012</td>
</tr>
</tbody>
</table>

- Max document length = 140 characters; single author per document; no post-editing

<table>
<thead>
<tr>
<th>Corpus</th>
<th>Documents</th>
<th>Average words per document</th>
</tr>
</thead>
<tbody>
<tr>
<td>Twitter-1</td>
<td>1,000,000</td>
<td>11.8 ± 8.3</td>
</tr>
<tr>
<td>Twitter-2</td>
<td>1,000,000</td>
<td>11.6 ± 8.1</td>
</tr>
</tbody>
</table>
**COMMENTS**

- All comments associate with YouTube videos in dataset of O’Callaghan et al. [2012]
- Max document length = 500 characters; single author per document; no post-editing

<table>
<thead>
<tr>
<th>Corpus</th>
<th>Documents</th>
<th>Average words per document</th>
</tr>
</thead>
<tbody>
<tr>
<td>COMMENTS</td>
<td>874,772</td>
<td>15.8±18.6</td>
</tr>
</tbody>
</table>
1M randomly-selected posts from top-1000 vBulletin-based forums in the Big Boards forum ranking

Max document length = site variable; single author per document; option for post-editing

<table>
<thead>
<tr>
<th>Corpus</th>
<th>Documents</th>
<th>Average words per document</th>
</tr>
</thead>
<tbody>
<tr>
<td>Forums</td>
<td>1,000,000</td>
<td>23.2±29.3</td>
</tr>
</tbody>
</table>
blogs

- 1M randomly-selected documents from Spinn3r dataset (ICWSM-2011 tier-one)
- Max document length = none; single author per document; post-editing possible

<table>
<thead>
<tr>
<th>Corpus</th>
<th>Documents</th>
<th>Average words per document</th>
</tr>
</thead>
<tbody>
<tr>
<td>Blogs</td>
<td>1,000,000</td>
<td>$147.7 \pm 339.3$</td>
</tr>
</tbody>
</table>
WIKIPEDIA

- 200K randomly-selected documents (≥ 500 bytes) from English Wikipedia
- Mediawiki markup removed with `wikidump`
- Max document length = none; multiple authors per document; post-editing possible

<table>
<thead>
<tr>
<th>Corpus</th>
<th>Documents</th>
<th>Average words per document</th>
</tr>
</thead>
<tbody>
<tr>
<td>WIKIPEDIA</td>
<td>200,000</td>
<td>281.2±363.8</td>
</tr>
</tbody>
</table>
BNC

- All documents in written portion of the British National Corpus
- Max document length = none; mostly single-author documents; post-editing possible

<table>
<thead>
<tr>
<th>Corpus</th>
<th>Documents</th>
<th>Average words per document</th>
</tr>
</thead>
<tbody>
<tr>
<td>BNC</td>
<td>3141</td>
<td>31609.0±30424.3</td>
</tr>
</tbody>
</table>
Clinical Note Corpora

Clinical note sources targeted in this research:

1. **VetCompass**: clinical notes from VetCompass UK
2. **NurseTriage**: Emergency Department nurse triage notes
3. **Radiology**: radiology clinical notes
VetCompass

- \( \approx 400M \) randomly-selected treatment notes from VetCompass UK
- Max document length = none; all single-author documents; post-editing not possible

<table>
<thead>
<tr>
<th>Corpus</th>
<th>Documents</th>
<th>Average words per document</th>
</tr>
</thead>
<tbody>
<tr>
<td>VetCompass</td>
<td>434,415</td>
<td>57.5±66.9</td>
</tr>
</tbody>
</table>
NURSE TRIAGE

- ≈56K nurse triage notes from the Emergency Department of Royal Melbourne Hospital [Kocbek et al., 2014]
- Max document length = none; all single-author documents; post-editing not possible

<table>
<thead>
<tr>
<th>Corpus</th>
<th>Documents</th>
<th>Average words per document</th>
</tr>
</thead>
<tbody>
<tr>
<td>NURSE TRIAGE</td>
<td>56837</td>
<td>24.9±14.1</td>
</tr>
</tbody>
</table>
Radiology

- \( \approx 9K \) radiology reports [Pestian et al., 2007]
- Max document length = none; all single-author documents; post-editing not possible

<table>
<thead>
<tr>
<th>Corpus</th>
<th>Documents</th>
<th>Average words per document</th>
</tr>
</thead>
<tbody>
<tr>
<td>Radiology</td>
<td>8825</td>
<td>108.0±52.8</td>
</tr>
</tbody>
</table>
Bringing it all Together ...

<table>
<thead>
<tr>
<th>Corpus</th>
<th>Documents</th>
<th>Average words per document</th>
</tr>
</thead>
<tbody>
<tr>
<td>Twitter-1</td>
<td>1,000,000</td>
<td>11.8±8.3</td>
</tr>
<tr>
<td>Twitter-2</td>
<td>1,000,000</td>
<td>11.6±8.1</td>
</tr>
<tr>
<td>Comments</td>
<td>874,772</td>
<td>15.8±18.6</td>
</tr>
<tr>
<td>Forums</td>
<td>1,000,000</td>
<td>23.2±29.3</td>
</tr>
<tr>
<td>Blogs</td>
<td>1,000,000</td>
<td>147.7±339.3</td>
</tr>
<tr>
<td>Wikipedia</td>
<td>200,000</td>
<td>281.2±363.8</td>
</tr>
<tr>
<td>BNC</td>
<td>3141</td>
<td>31609.0±30424.3</td>
</tr>
<tr>
<td>VetCompass</td>
<td>434,415</td>
<td>57.5±66.9</td>
</tr>
<tr>
<td>NurseTriage</td>
<td>56837</td>
<td>24.9±14.1</td>
</tr>
<tr>
<td>Radiology</td>
<td>8825</td>
<td>108.0±52.8</td>
</tr>
</tbody>
</table>
Corpus Preprocessing

- We apply the following preprocessing to all corpora (taken directly from Baldwin et al. [2013]):

  1. language identification with `langid.py`; non-English documents filtered from corpus
  2. sentence-tokenise with `tokenizer`, based on findings of Read et al. [2012]
  3. tokenise and POS tag with `TweetNLP 0.3`
  4. remove all “non-linguistic” tokens, on basis of `TweetNLP`

  ```
  @helloworld Swinging with the #besties! #awesome
  ↓
  @helloworld Swinging with the #besties! #awesome
  ```
Contents

1. The Eight “P”s of Clinical NLP
2. VetCompass
3. The Appeals/Challenges of Clinical NLP
4. Cross-comparison of Social Media and Clinical Text Sources
   - Corpora
   - Intra-Corpus Analysis
   - Inter-Corpus Analysis
   - Summary
5. Concluding Remarks
## Lexical Analysis

- Analysis of the average word and sentence length:

<table>
<thead>
<tr>
<th>Corpus</th>
<th>Word length</th>
<th>Sentence length</th>
</tr>
</thead>
<tbody>
<tr>
<td>Twitter-1</td>
<td>3.8±2.4</td>
<td>9.2±6.4</td>
</tr>
<tr>
<td>Twitter-2</td>
<td>3.8±2.4</td>
<td>9.0±6.3</td>
</tr>
<tr>
<td>Comments</td>
<td>3.9±3.2</td>
<td>10.5±10.1</td>
</tr>
<tr>
<td>Forums</td>
<td>3.8±2.3</td>
<td>14.2±12.7</td>
</tr>
<tr>
<td>Blogs</td>
<td>4.1±2.8</td>
<td>18.5±24.8</td>
</tr>
<tr>
<td>Wikipedia</td>
<td>4.5±2.8</td>
<td>21.9±16.2</td>
</tr>
<tr>
<td>BNC</td>
<td>4.3±2.8</td>
<td>19.8±14.5</td>
</tr>
<tr>
<td>VetCompass</td>
<td>4.2±2.9</td>
<td>8.9±8.1</td>
</tr>
<tr>
<td>NurseTriage</td>
<td>4.0±2.8</td>
<td>10.5±8.4</td>
</tr>
<tr>
<td>Radiology</td>
<td>4.8±3.3</td>
<td>10.3±6.1</td>
</tr>
</tbody>
</table>
## Lexical Analysis

- Analysis of the rate of out-of-vocabulary words (cf. aspell):

<table>
<thead>
<tr>
<th>Corpus</th>
<th>Word length</th>
<th>Sentence length</th>
<th>%OOV</th>
</tr>
</thead>
<tbody>
<tr>
<td>Twitter-1</td>
<td>3.8±2.4</td>
<td>9.2±6.4</td>
<td>24.6</td>
</tr>
<tr>
<td>Twitter-2</td>
<td>3.8±2.4</td>
<td>9.0±6.3</td>
<td>24.0</td>
</tr>
<tr>
<td>Comments</td>
<td>3.9±3.2</td>
<td>10.5±10.1</td>
<td>19.8</td>
</tr>
<tr>
<td>Forums</td>
<td>3.8±2.3</td>
<td>14.2±12.7</td>
<td>18.1</td>
</tr>
<tr>
<td>Blogs</td>
<td>4.1±2.8</td>
<td>18.5±24.8</td>
<td>20.6</td>
</tr>
<tr>
<td>Wikipedia</td>
<td>4.5±2.8</td>
<td>21.9±16.2</td>
<td>19.0</td>
</tr>
<tr>
<td>BNC</td>
<td>4.3±2.8</td>
<td>19.8±14.5</td>
<td>16.9</td>
</tr>
<tr>
<td>VetCompass</td>
<td>4.2±2.9</td>
<td>8.9±8.1</td>
<td>38.2</td>
</tr>
<tr>
<td>NurseTriage</td>
<td>4.0±2.8</td>
<td>10.5±8.4</td>
<td>41.9</td>
</tr>
<tr>
<td>Radiology</td>
<td>4.8±3.3</td>
<td>10.3±6.1</td>
<td>26.1</td>
</tr>
</tbody>
</table>
## Lexical Analysis

- Analysis of the rate of out-of-vocabulary words with lexical normalisation [Han et al., 2012]:

<table>
<thead>
<tr>
<th>Corpus</th>
<th>Word length (±)</th>
<th>Sentence length (±)</th>
<th>-norm %OOV</th>
<th>+norm %OOV</th>
</tr>
</thead>
<tbody>
<tr>
<td>Twitter-1</td>
<td>3.8±2.4</td>
<td>9.2±6.4</td>
<td>24.6</td>
<td>22.5</td>
</tr>
<tr>
<td>Twitter-2</td>
<td>3.8±2.4</td>
<td>9.0±6.3</td>
<td>24.0</td>
<td>22.2</td>
</tr>
<tr>
<td>Comments</td>
<td>3.9±3.2</td>
<td>10.5±10.1</td>
<td>19.8</td>
<td>18.4</td>
</tr>
<tr>
<td>Forums</td>
<td>3.8±2.3</td>
<td>14.2±12.7</td>
<td>18.1</td>
<td>17.1</td>
</tr>
<tr>
<td>Blogs</td>
<td>4.1±2.8</td>
<td>18.5±24.8</td>
<td>20.6</td>
<td>20.3</td>
</tr>
<tr>
<td>Wikipedia</td>
<td>4.5±2.8</td>
<td>21.9±16.2</td>
<td>19.0</td>
<td>18.8</td>
</tr>
<tr>
<td>BNC</td>
<td>4.3±2.8</td>
<td>19.8±14.5</td>
<td>16.9</td>
<td>16.8</td>
</tr>
<tr>
<td>VetCompass</td>
<td>4.2±2.9</td>
<td>8.9±8.1</td>
<td>38.2</td>
<td>37.2</td>
</tr>
<tr>
<td>NurseTriage</td>
<td>4.0±2.8</td>
<td>10.5±8.4</td>
<td>41.9</td>
<td>40.1</td>
</tr>
<tr>
<td>Radiology</td>
<td>4.8±3.3</td>
<td>10.3±6.1</td>
<td>26.1</td>
<td>26.0</td>
</tr>
</tbody>
</table>
Lexical Analysis: Overall Findings

- Slight difference in average word length (esp. for Radiology); much larger difference in average sentence length: \{Wikipedia, BNC, Blogs\} > Forums > \{Comments, Twitter, VetCompass, NurseTriage, Radiology\}

- With OOV%, VetCompass and NurseTriage much higher than social media corpora, with Radiology comparable to Twitter-1/2; OOV method trained on social media largely ineffectual over clinical text
Contents

1 The Eight “P”s of Clinical NLP
2 VetCompass
3 The Appeals/Challenges of Clinical NLP
4 Cross-comparison of Social Media and Clinical Text Sources
   - Corpora
   - Intra-Corpus Analysis
   - Inter-Corpus Analysis
   - Summary
5 Concluding Remarks
Corpus Similarity

- Next, we compare the (original) corpora with one another based on trigram language model-based perplexity (SRILM, with Kneser-Ney smoothing).
- For each corpus, split into $5 \times 1M$ sentence sub-corpora, train a LM over 4 sub-corpora, and apply it to a 1M sub-corpus from another domain, or the heldout sub-corpus for that domain; average the perplexity scores.
Trigram Perplexity Results I

<table>
<thead>
<tr>
<th>TWITTER-1</th>
<th>COMMENTS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Twitter-1</td>
<td>191</td>
</tr>
<tr>
<td>Twitter-2</td>
<td>742</td>
</tr>
<tr>
<td>Comments</td>
<td>576</td>
</tr>
<tr>
<td>Blogs</td>
<td>1,048</td>
</tr>
<tr>
<td>Forums</td>
<td>593</td>
</tr>
<tr>
<td>Wikipedia</td>
<td>2,187</td>
</tr>
<tr>
<td>BNC</td>
<td>1,555</td>
</tr>
<tr>
<td>VetCompass</td>
<td>5,247</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>TWITTER-2</th>
<th>BLOGS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Twitter-1</td>
<td>1,305</td>
</tr>
<tr>
<td>Twitter-2</td>
<td>1,359</td>
</tr>
<tr>
<td>Comments</td>
<td>112</td>
</tr>
<tr>
<td>Blogs</td>
<td>1,061</td>
</tr>
<tr>
<td>Forums</td>
<td>488</td>
</tr>
<tr>
<td>Wikipedia</td>
<td>1,903</td>
</tr>
<tr>
<td>BNC</td>
<td>1,437</td>
</tr>
<tr>
<td>VetCompass</td>
<td>5,889</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>TWITTER-2</th>
<th>BLOGS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Twitter-1</td>
<td>1,532</td>
</tr>
<tr>
<td>Twitter-2</td>
<td>1,566</td>
</tr>
<tr>
<td>Comments</td>
<td>174</td>
</tr>
<tr>
<td>Blogs</td>
<td>697</td>
</tr>
<tr>
<td>Forums</td>
<td>434</td>
</tr>
<tr>
<td>Wikipedia</td>
<td>805</td>
</tr>
<tr>
<td>BNC</td>
<td>616</td>
</tr>
<tr>
<td>VetCompass</td>
<td>4,201</td>
</tr>
</tbody>
</table>
Trigram Perplexity Results II

**FORUMS**

<table>
<thead>
<tr>
<th></th>
<th>VetCompass</th>
<th>BNC</th>
<th>Wikipedia</th>
<th>Forums</th>
<th>Blogs</th>
<th>Twitter-1</th>
<th>Twitter-2</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>Twitter-1</td>
<td>1,451</td>
<td>3,488</td>
<td>4,960</td>
<td>1,984</td>
<td>1,089</td>
<td>4,548</td>
<td>2,370</td>
<td>158</td>
</tr>
<tr>
<td>Twitter-2</td>
<td>1,482</td>
<td>3,612</td>
<td>4,955</td>
<td>1,984</td>
<td>1,043</td>
<td>4,548</td>
<td>3,727</td>
<td>158</td>
</tr>
<tr>
<td>Comments</td>
<td>526</td>
<td>1,564</td>
<td>1,984</td>
<td>1,089</td>
<td>1,043</td>
<td>4,548</td>
<td>2,370</td>
<td>158</td>
</tr>
<tr>
<td>Blogs</td>
<td>740</td>
<td>894</td>
<td>1,089</td>
<td>1,043</td>
<td>1,043</td>
<td>3,727</td>
<td>3,489</td>
<td>145</td>
</tr>
<tr>
<td>Forums</td>
<td>151</td>
<td>784</td>
<td>1,089</td>
<td>1,043</td>
<td>1,043</td>
<td>3,727</td>
<td>3,489</td>
<td>145</td>
</tr>
<tr>
<td>Wikipedia</td>
<td>1,215</td>
<td>722</td>
<td>1,089</td>
<td>1,043</td>
<td>1,043</td>
<td>3,727</td>
<td>3,489</td>
<td>145</td>
</tr>
<tr>
<td>BNC</td>
<td>975</td>
<td>162</td>
<td>1,089</td>
<td>1,043</td>
<td>1,043</td>
<td>3,727</td>
<td>3,489</td>
<td>145</td>
</tr>
</tbody>
</table>

**WIKIPEDIA**

<table>
<thead>
<tr>
<th></th>
<th>VetCompass</th>
<th>BNC</th>
<th>Wikipedia</th>
<th>Forums</th>
<th>Blogs</th>
<th>Twitter-1</th>
<th>Twitter-2</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>Twitter-1</td>
<td>4,694</td>
<td>6,080</td>
<td>9,124</td>
<td>162</td>
<td>595</td>
<td>7,852</td>
<td>8,106</td>
<td>158</td>
</tr>
<tr>
<td>Twitter-2</td>
<td>4,964</td>
<td>6,080</td>
<td>9,124</td>
<td>162</td>
<td>595</td>
<td>7,852</td>
<td>8,106</td>
<td>158</td>
</tr>
<tr>
<td>Comments</td>
<td>975</td>
<td>722</td>
<td>1,089</td>
<td>1,043</td>
<td>1,043</td>
<td>3,727</td>
<td>3,489</td>
<td>145</td>
</tr>
<tr>
<td>Blogs</td>
<td>975</td>
<td>722</td>
<td>1,089</td>
<td>1,043</td>
<td>1,043</td>
<td>3,727</td>
<td>3,489</td>
<td>145</td>
</tr>
<tr>
<td>Forums</td>
<td>151</td>
<td>784</td>
<td>1,089</td>
<td>1,043</td>
<td>1,043</td>
<td>3,727</td>
<td>3,489</td>
<td>145</td>
</tr>
<tr>
<td>Wikipedia</td>
<td>1,215</td>
<td>722</td>
<td>1,089</td>
<td>1,043</td>
<td>1,043</td>
<td>3,727</td>
<td>3,489</td>
<td>145</td>
</tr>
<tr>
<td>BNC</td>
<td>975</td>
<td>162</td>
<td>1,089</td>
<td>1,043</td>
<td>1,043</td>
<td>3,727</td>
<td>3,489</td>
<td>145</td>
</tr>
</tbody>
</table>

Perplexity
Corpus Similarity Findings

- Overall, there appears to be a partial ordering in corpus similarity: Twitter-1/2 ≡ Comments < Forums < Blogs < BNC < Wikipedia
- The social media sources which are most similar to VetCompass is Blogs and Forums, but big disparity over other social media sources
- Despite this, the in-domain perplexity for VetCompass is lower than all of the social media sources, suggesting that it is internally highly homogeneous
Other Analysis I Wanted to do …

- Would have liked to have done perplexity analysis over the other two clinical text sources, but not enough data
- Syntactic analysis would be a natural next step [Friedman et al., 2002, Stetson et al., 2002, Baldwin et al., 2013]
- More formally exploring formal “domain” differences between data sources to better quantify the differences between clinical domains, and between clinical and social media text would also be interesting to explore [Lippincott et al., 2010, Verspoor et al., 2009, McClosky et al., 2010, MacKinlay et al., 2010]
Contents

1. The Eight “P”s of Clinical NLP
2. VetCompass
3. The Appeals/Challenges of Clinical NLP
4. Cross-comparison of Social Media and Clinical Text Sources
   - Corpora
   - Intra-Corpus Analysis
   - Inter-Corpus Analysis
   - Summary
5. Concluding Remarks
Summary of Overall Findings I

- Large-scale analysis of a range of social media vs. clinical text sources (+ BNC), in terms of lexical analysis and corpus similarity
Summary of Overall Findings I

- Large-scale analysis of a range of social media vs. clinical text sources (+ BNC), in terms of lexical analysis and corpus similarity
- How “noisy” are social media and clinical text sources, and in what ways?
Summary of Overall Findings I

- Large-scale analysis of a range of social media vs. clinical text sources (+ BNC), in terms of lexical analysis and corpus similarity
- How “noisy” are social media and clinical text sources, and in what ways?
  - much more lexical “noise” in the clinical text sources (esp. VetCompass, NurseTriage), and POS tagging + lexical normalisation (training on social media text) does little to reduce it
Summary of Overall Findings II

- Are different social media and clinical text sources equally “noisy”?
Summary of Overall Findings II

Are different social media and clinical text sources equally “noisy”?

- no; Twitter definitely on the noisy end of the spectrum, and Wikipedia “cleaner” than BNC, but the spread is less than you might think; strong suspicion that VetCompass and NurseTriage will come out “noisier” than Twitter-1/2, but watch this space
Summary of Overall Findings II

- Are different social media and clinical text sources equally “noisy”?
  - no; Twitter definitely on the noisy end of the spectrum, and Wikipedia “cleaner” than BNC, but the spread is less than you might think; strong suspicion that VetCompass and NurseTriage will come out “noisier” than Twitter-1/2, but watch this space

- Ultimately, are the differences between social media and clinical text sources all that great?
Summary of Overall Findings II

- Are different social media and clinical text sources equally “noisy”?
  - no; Twitter definitely on the noisy end of the spectrum, and Wikipedia “cleaner” than BNC, but the spread is less than you might think; strong suspicion that VetCompass and NurseTriage will come out “noisier” than Twitter-1/2, but watch this space

- Ultimately, are the differences between social media and clinical text sources all that great?
  - yes, clinical text is a very different beast to social media text; different social media sources are much closer to one another than they are to clinical text
Talk Outline

1. The Eight “P”s of Clinical NLP
2. VetCompass
3. The Appeals/Challenges of Clinical NLP
4. Cross-comparison of Social Media and Clinical Text Sources
   - Corpora
   - Intra-Corpus Analysis
   - Inter-Corpus Analysis
   - Summary
5. Concluding Remarks
Final Words

- Clinical text is a treasure trove of knowledge/information, but equally a Pandora’s box of data access and NLP challenges.
- **VetCompass** potentially offers insights into clinical text at scale, with lots of cool work to be done.
Acknowledgements

- Thanks to the inimitable Karin Verspoor for valuable assistance with data and sagacious advice on an earlier version of this presentation, and Noel Kennedy, Paul McGreevy and VetCompass for the VetCompass data sample.

- This research was funded in part by the Australian Research Council.
References


References II


References III


Karin Verspoor, Kevin B. Cohen, and Lawrence Hunter. The textual characteristics of traditional and open access scientific journals are similar. *BMC Bioinformatics*, 10(1), 2009.