VetCompass: Clinical Natural Language Processing for Animal Health

Timothy Baldwin

THE UNIVERSITY OF MELBOURNE

Talk Outline

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

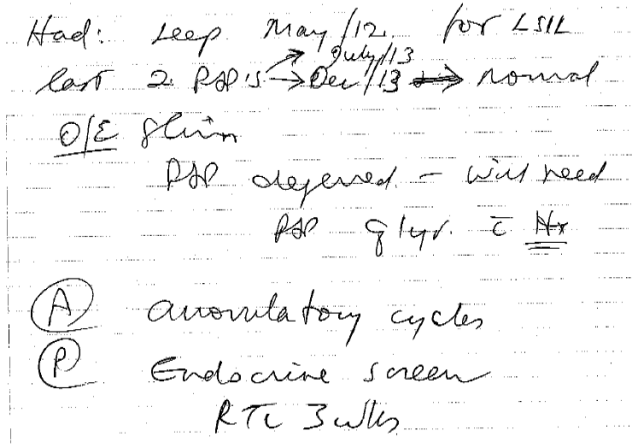
  ... for clinical NLP, we have the Eight "P"s ...

"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, ...

o 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

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

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?

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?

  ... 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:

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?

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

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
  5. WIKIPEDIA: documents from English Wikipedia
- 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:
  A. Twitter-1 = 22 Sep, 2011
  B. Twitter-2 = 22 Feb, 2012
- Max document length = 140 characters; single author per document; no post-editing

Corpus	Documents	Average words per document
Twitter-1	1,000,000	11.8±8.3
Twitter-2	1,000,000	11.6±8.1

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

Corpus	Documents	Average words per document
Comments	874,772	15.8±18.6

Forums

- 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

Corpus	Documents	Average words per document
Forums	1,000,000	23.2±29.3

Blogs

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

Corpus	Documents	Average words per document
Blogs	1,000,000	147.7±339.3

Wikipedia

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

Corpus	Documents	Average words per document
Wikipedia	200,000	281.2±363.8

BNC

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

Corpus	Documents	Average words per document
BNC	3141	31609.0±30424.3

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

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

Corpus	Documents	Average words per document
VetCompass	434,415	57.5±66.9

NurseTriage

- ≈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

Corpus	Documents	Average words per document
NurseTriage	56837	24.9±14.1

Radiology

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

Corpus	Documents	Average words per document
Radiology	8825	108.0±52.8

Bringing it all Together ...

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

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

Lexical Analysis

- Analysis of the average word and sentence length:

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

Lexical Analysis

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

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

Lexical Analysis

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

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

Lexical Analysis: Overall Findings

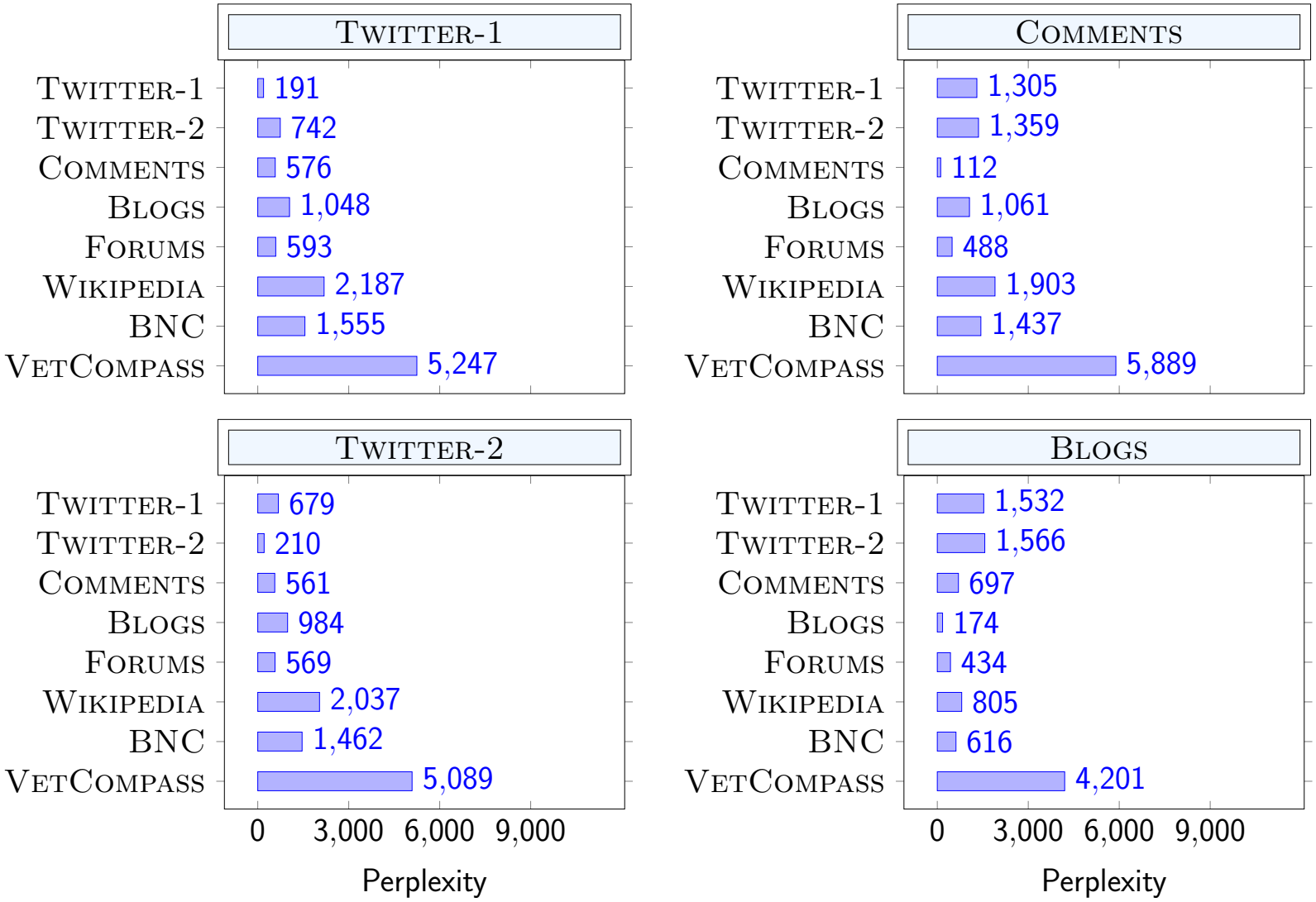
- Slight difference in average word length (esp. for Radiology); much larger difference in average sentence length: {Wikipedia,BNC,Blogs } > Forums > {Comments,Twitter,VetCompass,NurseTriage, Radiology }
- With OOV%, VetCompass and NurseTriage much higher than social media corpora, with Radiology comparable to Twitter-1/2; OOV method trained on social media largely ineffectual over clinical text

Contents

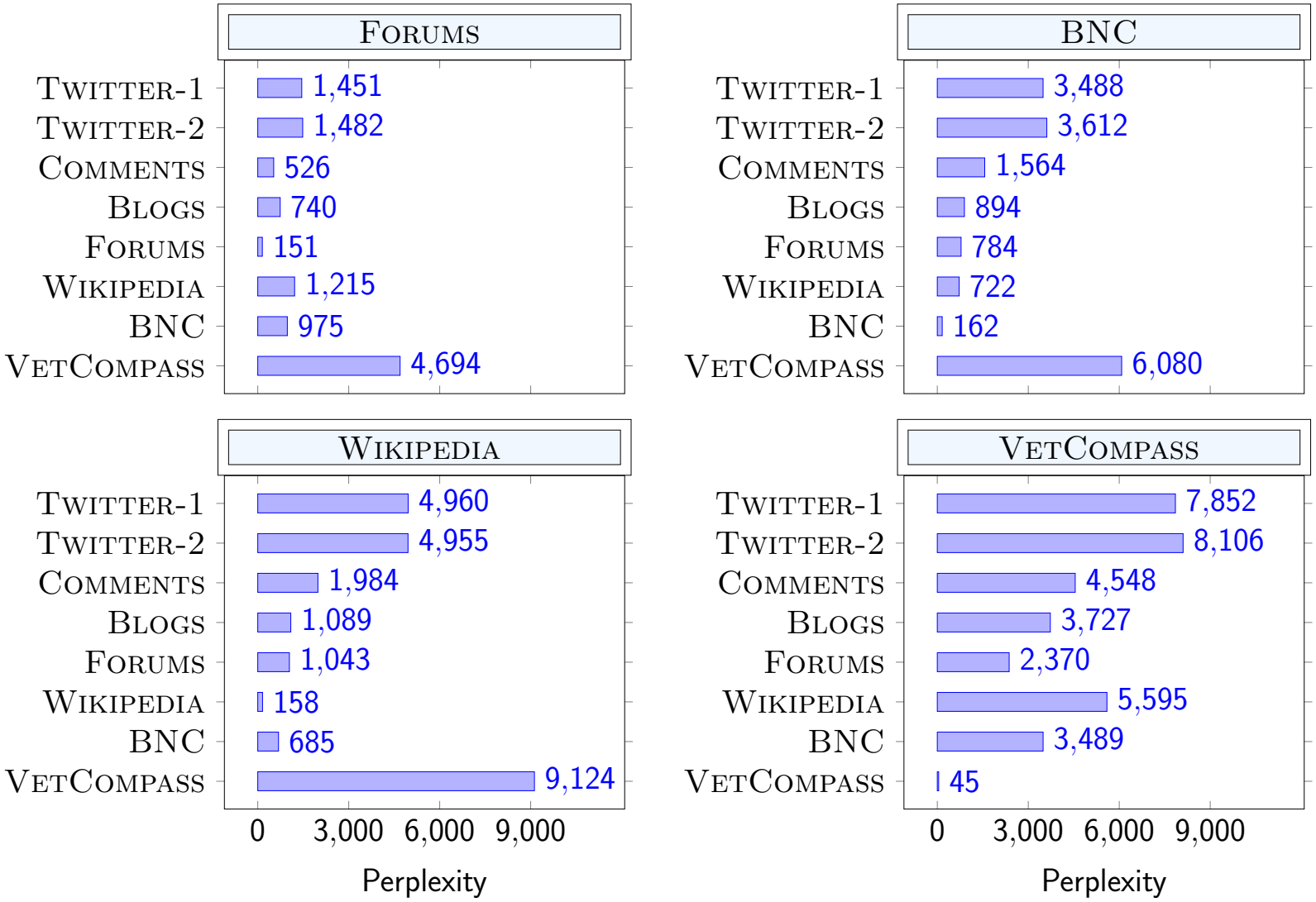
Corpus Similarity

- Next, we compare the (original) corpora with one another based on trigram language model-based perplexity (SRILM, with Knesser-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



Trigram Perplexity Results II



Corpus Similarity Findings

- Overall, there appears to be a partial ordering in corpus similarity: $\textsc{Twitter}\text{-}1/2 \equiv \textsc{Comments} < \textsc{Forums} < \textsc{Blogs} < \textsc{BNC} < \textsc{Wikipedia}$
- The social media sources which are most similar to $\textsc{VetCompass}$ is $\textsc{Blogs}$ and $\textsc{Forums}$, but big disparity over other social media sources
- Despite this, the in-domain perplexity for $\textsc{VetCompass}$ 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

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

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 I

Timothy Baldwin, Paul Cook, Marco Lui, Andrew MacKinlay, and Li Wang. How noisy social media text, how diffrnt social media sources? In *Proceedings of the 6th International Joint Conference on Natural Language Processing (IJCNLP 2013)*, pages 356–364, Nagoya, Japan, 2013.

Carol Friedman, Pauline Kra, and Andrey Rzhetsky. Two biomedical sublanguages: a description based on the theories of Zellig Harris. *Journal of Biomedical Informatics*, 35(4):222–235, 2002.

Bo Han, Paul Cook, and Timothy Baldwin. Automatically constructing a normalisation dictionary for microblogs. In *Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning 2012 (EMNLP-CoNLL 2012)*, pages 421–432, Jeju, Korea, 2012.

Simon Kocbek, Karin Verspoor, and Wray Buntine. Exploring temporal patterns in emergency department triage notes with topic models. In *Proceedings of the Australasian Language Technology Workshop 2014 (ALTW 2014)*, pages 113–117, 2014.

Tom Lippincott, Diarmuid Ó Séaghdha, Lin Sun, and Anna Korhonen. Exploring variation across biomedical subdomains. In *Proceedings of the 23rd International Conference on Computational Linguistics (COLING 2010)*, pages 689–697, 2010.

References II

Andrew MacKinlay, Timothy Baldwin, Dan Flickinger, and Rebecca Dridan. Cross-domain effects on parse selection for precision grammars. *Journal of Research on Language and Computation*, 8(4):299–340, 2010.

Andrew McAfee and Erik Brynjolfsson. Big data: The management revolution. *Harvard Business Review*, 59:60–69, 2012.

David McClosky, Eugene Charniak, and Mark Johnson. Automatic domain adaptation for parsing. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 28–36, 2010.

E. Nygren and P. Henriksson. Reading the medical record I. *Computer Methods and Programs in Biomedicine*, 39:1–12, 1992.

Derek O'Callaghan, Martin Harrigan, Joe Carthy, and Pádraig Cunningham. Network analysis of recurring YouTube spam campaigns. In *Proceedings of the 6th International Conference on Weblogs and Social Media (ICWSM 2012)*, pages 531–534, Dublin, Ireland, 2012.

P. John Pestian, Chris Brew, Pawel Matykiewicz, D.J. Hovermale, Neil Johnson, Bretonnel K. Cohen, and Wlodzislaw Duch. A shared task involving multi-label classification of clinical free text. In *Biological, translational, and clinical language processing*, pages 97–104, Prague, Czech Republic, 2007.

References III

Lise Poissant, Jennifer Pereira, Robyn Tamblyn, and Yuko Kawasumi. The impact of electronic health records on time efficiency of physicians and nurses: a systematic review. *Journal of the American Medical Informatics Association*, 12(5):505–516, 2005.

Jonathon Read, Rebecca Dridan, Stephan Oepen, and Lars Jørgen Solberg. Sentence boundary detection: A long solved problem? In *Proceedings of COLING 2012: Posters*, pages 985–994, Mumbai, India, 2012. URL `http://www.aclweb.org/anthology/C12-2096`.

Peter D Stetson, Stephen B Johnson, Matthew Scotch, and George Hripcsak. The sublanguage of cross-coverage. In *Proceedings of the AMIA Symposium*, pages 742–746, 2002.

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.