Impact of Corpus Diversity and Complexity on NER Performance

Tatyana Shmanina[1,2], **Ingrid Zukerman**[1,2], **Antonio Jimeno Yepes**[1,3], **Lawrence Cavedon**[1,3], **Karin Verspoor**[1,3]
[1]NICTA Victoria Research Laboratory, Melbourne, Australia
[2]Clayton School of Information Technology, Monash University, Australia
[3]Department of Computing and Information Systems, University of Melbourne, Australia
Tatyana.Shmanina@nicta.com.au, Ingrid.Zukerman@monash.edu,
Antonio.Jimeno@gmail.com,{Lawrence.Cavedon,Karin.Verspoor}@nicta.com.au

Abstract

We describe a cross-corpora evaluation of disease mention recognition for two annotated biomedical corpora: the *Human Variome Project Corpus* and the *Arizona Disease Corpus*. Our analysis of the performance of a state-of-the-art NER tool in terms of the characteristics and annotation schema of these corpora shows that these factors significantly affect performance.

1 Introduction

The recent growth of on-line biomedical literature has spawned an increasing number of NLP tools for content analysis that help researchers and practitioners access the latest developments in their fields. Examples of these tools include: *BANNER* – a *Named Entity Recognizer* (*NER*) for the biomedical domain (Leaman and Gonzalez, 2008); *ABNER* – a NER for molecular biology (Settles, 2004); and *Whatizit* – a Web service which provides functionality to perform text-mining tasks (Rebholz-Schuhmann et al., 2008). These tools in turn require the development of annotated training corpora, e.g., (Kim et al., 2003; Rosario and Hearst, 2004; Kulick et al., 2004; Pestian et al., 2007; Jimeno-Yepes et al., 2008; Bada et al., 2012).

Studies have been conducted to examine the performance of different NLP tools on a single corpus, e.g., (Jacob et al., 2013; Verspoor et al., 2012). However, experience shows that the characteristics of a corpus influence performance, e.g., (Cao and Zukerman, 2012) for sentiment analysis and (Pyysalo et al., 2008) in the biomedical space. In this paper, we analyze how the characteristics and annotation schemas of two corpora influence *BANNER*'s performance on the recognition of diseases (note that *BANNER* outperforms *ABNER* in the recognition of diseases (Leaman and Gonzalez, 2008)). The corpora in question are the *Human Variome Project Corpus* (*HVPC*) developed at NICTA (Verspoor et al., 2013), and the *Arizona Disease Corpus* (*AZDC*) – a popular medical resource developed at the University of Arizona (Leaman et al., 2009).[1]

Our results show that *BANNER*'s performance on *HVPC* significantly exceeds its performance on *AZDC*. This is (at least partly) explained by differences in corpus characteristics, such as reduced disease mention diversity resulting from *HVPC*'s specific focus, and by some requirements of *HVPC*'s annotation schema. These observations suggest that corpus analysis should be conducted along with performance evaluation in order to appropriately assess the obtained results and the suitability of a corpus for training general NER tools.

2 Biomedical Corpora

AZDC is a biomedical textual resource focusing on disease annotation (Leaman et al., 2009). It was extracted from a corpus created by Craven and Kumlien (1999), which consists of sentences selected from MEDLINE® abstracts via queries for six proteins. All disease mentions in *AZDC* are annotated, with each disease annotation containing a *Unified Medical Language System®* (*UMLS®*) concept unique identifier (where possible).

HVPC is an annotated biomedical textual resource pertaining to human genetic variation and its relation to diseases (Verspoor et al., 2013). At present, the corpus comprises ten double-annotated plain-text full journal publications on inherited colorectal cancer, which were selected on the basis of their relevance to the genetics of the Lynch Syndrome. The annotation schema, which is tailored to the focus of the corpus, covers thirteen *relations*, such as "gene-has-mutation", "mutation-has-size" and "disease-related-to-body-part"; and eleven *entity types*, such as genomic categories (e.g., "gene", "mutation"), phenotypic categories (e.g., "disease", "body-part"), categories

[1]Of the above corpora, only Kulick *et al.*'s focuses on diseases at the same level of detail as the corpora considered in this paper, and may be investigated in the future.

Tatyana Shmanina, Ingrid Zukerman, Antonio Jimeno Yepes, Lawrence Cavedon and Karin Verspoor. 2013. Impact of Corpus Diversity and Complexity on NER Performance. In *Proceedings of Australasian Language Technology Association Workshop*, pages 91−95.

related to the occurrence of mutations in a disease (e.g., "age", "ethnicity"), and a "characteristic" category as a catch-all for information of interest that is otherwise uncategorized.

2.1 Comparison of Annotation Schemas

Both *HVPC* and *AZDC* annotate duplicate disease mentions in the same sentence, and abbreviations specific to the analyzed article (e.g., "Huntington disease (HD)"). In addition, they do not annotate stand-alone generic words (e.g., "disease", "syndrome"), and disease names embedded into entities of other types (e.g., "Peter MacCallum *Cancer* Centre"). However, there are significant differences between these annotation schemas:

- *HVPC*'s annotation guidelines define the "disease" entity type as "an abnormal condition affecting the body of an organism", and annotates modifiers such "healthy", "unaffected" and "normal" as diseases of healthy individuals. In contrast, *AZDC* requires that disease mentions correspond to one of the several semantic types of the UMLS® Semantic Group "disorders" (e.g., "disease or syndrome", "injury or poisoning", "mental dysfunction", "sign or symptom"). As a result, disease effects are annotated as diseases in *AZDC*, but not in *HVPC*.
- *AZDC* requires mention boundaries to be set to a minimum span of text necessary to describe the most specific form of a disease. In contrast, *HVPC* seems to be more restrictive with respect to disease mention boundaries. Specifically, many of the modifiers describing the type of a disease (which are included in disease mentions in *AZDC*) are attributed to the "characteristic" entity type (Section 1). For example, "classical galactosemia" and "unilateral retinoblastoma" are disease mentions according to *AZDC*, while only the head noun is a disease mention according to *HVPC*.
- *HVPC* annotates only the last and most complete part of a disease coordination[2] (e.g., in "breast and *ovarian cancer*", "breast" is annotated as a body part[3]), while *AZDC* annotates a coordination as separate but overlapping mentions of a disease (e.g., "*breast and ovarian cancer*" and "*ovarian cancer*").

[2]This was originally done in response to the BRAT annotation tool (Stenetorp et al., 2012) not allowing annotation of discontinuous entities (since rectified).

[3]A refinement is to consider *(body-part, disease)* related pairs as multi-word disease names, which would boost the mention-length counts for *HVPC* in Figure 2.

These aspects account for the simplicity, brevity and higher structural regularity of *HVPC* disease mentions compared to those in *AZDC* (Section 2.2).

2.2 Comparison of Corpora Parameters

We have analyzed *HVPC* and *AZDC* with respect to the following parameters: size of the corpora in terms of number of sentences and tokens; number of disease mentions and unique disease mentions; and distribution of sentence length, disease mention length and disease mention frequency. The results, which appear in Tables 1 and 2, and Figures 1 and 2, reveal the following differences between *HVPC* and *AZDC*, which explain why *AZDC* is more difficult to analyze automatically than *HVPC*:

- **Unique disease mentions –** The ratio of *unique* disease mentions to *total* disease mentions in *HVPC* (8.4%) is much lower than in *AZDC* (37.2%) (Table 1). In addition, in *HVPC* a small set of unique mentions has very high frequency compared to *AZDC* (Table 2). These properties of *HVPC* may be attributed to its narrow focus on the Lynch Syndrome.
- **Sentence length and complexity –** In general, sentence length is significantly higher in *AZDC* (Figure 1). This may be attributed in part to the way in which *HVPC* and *AZDC* were constructed: *AZDC* contains only sentences extracted from biomedical paper abstracts, while *HVPC* consists of full papers, which in addition to sentences, contain section headings and table and figure captions.
- **Disease mention length –** Most disease mentions in *HVPC* consist of 1 or 2 terms, while *AZDC* contains a large number of multi-word complex disease mentions (Figure 2).

3 NER Performance

In this section, we describe the experiments we performed to evaluate the performance achieved for *HVPC* and *AZDC* by a state-of-the-art NER tool, viz *BANNER* (Leaman and Gonzalez, 2008) (Section 1). We also analyze the types of errors made by *BANNER* on each corpus, and discuss their connection to the annotation guidelines.

BANNER is a NER system developed for use in the biomedical domain. It uses a mechanism based on Conditional Random Fields (CRF) (Lafferty et al., 2001) to assign labels to input tokens, and considers the following features: (1) lemma for a token; (2) part of speech; (3) orthographic features, such as capitalization, presence of digits, prefixes and suffixes, and 2 and 3-character n-grams.

Parameter	*HVPC*	*AZDC*
# of sentences	2116	2783
# of tokens	52454	79950
Total # of disease mentions	1552	3228
# of unique disease mentions	130	1202

Table 1: Various quantitative parameters of *HVPC* and *AZDC*. Unique mentions refer to all (case-sensitive) textually identical disease mentions.

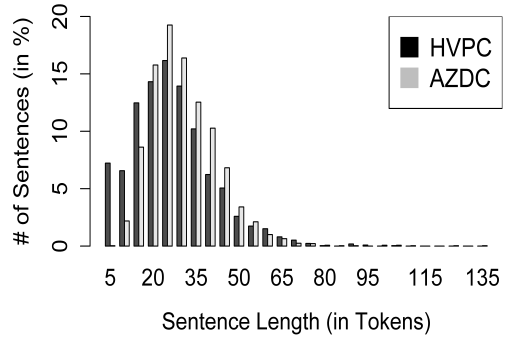

Figure 1: Distribution of sentence lengths (binned with step 5) in *HVPC* and *AZDC*.

3.1 Experimental Set-Up

***BANNER* configuration:** We used the 19th SVN revision of *BANNER* (`sourceforge.net/p/banner/code/HEAD/tree/`) with the following parameters: (1) parenthesis post-processing, post processing of abbreviations specific to an article, and numeric normalization switched "on"; (2) IOB (Inside, Outside, Begin) label model with second order CRF model; and (3) no dictionary.

Matching schemes: *BANNER*'s performance was assessed using the following matching schemes: (1) exact, (2) left border, (3) right border, (4) left or right border, (5) entity inclusion (one entity is a subset of another), and (6) entity overlap. The first scheme provides the most stringent measure of performance, while the other schemes provide different types of fuzzy matches.

Dataset preparation: *BANNER* contains a dataset loader specifically created for *AZDC*, while *HVPC* had to be segmented into sentences. This was done by training the OpenNLP sentence splitter (`opennlp.apache.org/`) on 70% of *HVPC*, and manually fixing the nine errors that remained after automatic sentence splitting.

3.2 Performance Evaluation

We performed 10-fold cross validation over both corpora, and employed the standard performance metrics of *Precision*, *Recall* and *F-score*.

Parameter	*HVPC*	*AZDC*
Frequency mean	11.94	2.73
Frequency standard deviation	22.39	5.65
Ratio of top N frequent mentions to all mentions		
$N = 10$	0.51	0.14
$N = 20$	0.73	0.22
$N = 30$	0.85	0.28

Table 2: Frequencies of disease mentions.

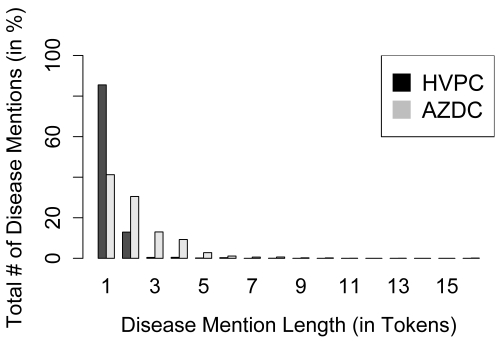

Figure 2: Distribution of disease mention lengths in *HVPC* and *AZDC*.

The results in Table 3 show that *BANNER* achieves excellent performance (*F-score*=0.9164) for *HVPC* on exact matches, which cannot be substantially improved by relaxing the matching scheme. In contrast, *AZDC*'s *F-score*=0.7365 for the exact matching scheme increases by up to 15% with matching-scheme relaxation.[4]

The good performance of *BANNER* on *HVPC* may be attributed to the single-disease focus of the corpus, its sentence brevity, and its disease-mention properties, which in turn are influenced by the annotation schema (Section 2). The latter may also explain the relative insensitivity of *BANNER* to the matching scheme relaxation: *BANNER* tends to have NE boundary detection problems mostly for long disease mentions, which are under-represented in *HVPC*.

With regard to *AZDC*, the results in Table 3 indicate that the main cause of the relatively low performance of *BANNER* is its inaccurate left boundary detection, which affects performance for both the exact and left-border schemes.

[4]In another set of experiments, *BANNER* trained on *AZDC* and tested on *HVPC* exhibited inferior performance (*F-score*=0.4453 for the exact matching scheme and *F-score*=0.6166 for overlap matching), thus confirming the large difference and non-interchangeability of these two datasets and their annotation schemas.

Scheme	Corpus	*Precision*	*Recall*	*F-score*
Exact	*AZDC*	0.7772	0.7003	0.7365
	HVPC	0.9322	0.9026	0.9164
Left	*AZDC*	0.8009	0.7217	0.7590
Border	*HVPC*	0.9372	0.9076	0.9214
Right	*AZDC*	0.8706	0.7844	0.8250
Border	*HVPC*	0.9512	0.9215	0.9353
Left or Right	*AZDC*	0.8870	0.7992	0.8406
Border	*HVPC*	0.9555	0.9258	0.9396
Inclusion	*AZDC*	0.8897	0.8016	0.8431
	HVPC	0.9593	0.9293	0.9433
Overlap	*AZDC*	0.8931	0.8046	0.8463
	HVPC	0.9599	0.9299	0.9439

Table 3: 10-fold X-validation for *AZDC* and *HVPC*.

3.3 NER Errors

Below we consider the errors identified in (Leaman et al., 2009) for *AZDC* (items 1-3), and add another type of error (item 4):

1. Improper handling of coordinations (*AZDC*), which occurs quite often, despite the addition of a coordination-handling post-processing step. *BANNER* tends to combine separate mentions of the form "disease1 and disease2" (false positives), while sometimes missing annotated coordinations (false negatives).
2. Inability to correctly detect boundaries of disease mentions (*AZDC* and *HVPC*). This problem is exacerbated in *AZDC* when diseases are referred to by their effects rather than their names (e.g., "premature periodontal destruction"), or disease names contain attributes (e.g., "high myopia").
3. Incorrect identification of acronyms and abbreviations specific to the analyzed article (*AZDC* and *HVPC*).
4. Overlooking (false negatives) or mistaken annotation (false positives) of disease names (*AZDC* and *HVPC*). In particular, this happens for words characterizing a health condition, e.g., "affected", "normal" or "healthy" (*HVPC* only), and diseases referred to by their effects (*AZDC* only).

This analysis confirms that the difference in *BANNER*'s performance on *HVPC* and *AZDC* is partly caused by differences in the annotation guidelines for these two corpora:

- *AZDC* contains many coordinations, while *HVPC*'s annotation guidelines circumvent the "coordination problem" (Section 2).
- Disease effects and characteristics are not annotated as disease names in *HVPC*. In contrast, the number of such mentions in *AZDC* is high, and its disease mentions in general are usually longer and more diverse than disease mentions in *HVPC*.

These factors explain the increased difficulty of disease-mention identification and mention-boundary detection in *AZDC* compared to *HVPC*.

3.4 Baseline Performance on *HVPC*

The simplicity of *HVPC* is further demonstrated by evaluating the performance of a very simple baseline algorithm that extracts disease mentions from *HVPC*. This algorithm applies the Unix string-matching utility *grep* to each word in a small (42 word) dictionary that was quickly constructed. The dictionary was created by collecting all the disease mentions and their morphological variations from the Wikipedia article about the Lynch Syndrome, and adding six terms ("healthy", "normal", "unaffected", "polyp", "polyps" and "polyposis"). The results obtained by this baseline for the exact matching scheme (*Precision* = 0.8777, *Recall* = 0.7352 and *F-score*=0.8001) are significantly better than the *BANNER* scores for *AZDC*.

4 Conclusion

In this paper, we presented a case study of two corpora with disease annotations. Our results show that the domain and construction method of a corpus, the restrictions imposed on disease definitions, and other annotation schema requirements are likely to have a high impact on NER performance. In particular, *HVPC* is an easy corpus for NER in comparison with *AZDC* due to its low lexical variability, the brevity and high regularity of its disease names, and the requirements of the *HVPC* annotation schema.

We conclude that corpus features identified in this paper are predictive of NER performance, and possibly of performance in other tasks, and should be taken into account during corpus selection. In particular, we note that *HVPC* is not very suitable for the development of NER tools for disease name recognition in general. However, this corpus may be useful for the development and assessment of (disease) relation extraction (RE) tools, as it minimizes the noise introduced by incorrect NER. In addition, it may be suitable for training NER and RE tools for applications focused on particular diseases.

Future research directions include studying other biomedical corpora and specializing high-diversity corpora (e.g., *AZDC*) to determine characteristics that most affect NER performance.

Acknowledgements

NICTA is funded by the Australian Government through the Department of Communications and by the Australian Research Council through the ICT Centre of Excellence Program.

References

M. Bada, M. Eckert, D. Evans, K. Garcia, K. Shipley, D. Sitnikov, W.A. Baumgartner, K. Bretonnel Cohen, K.M. Verspoor, J.A. Blake, and L.E. Hunter. 2012. Concept annotation in the CRAFT corpus. *BMC Bioinformatics*, 13(161).

M.D. Cao and I. Zukerman. 2012. Experimental evaluation of a lexicon- and corpus-based ensemble for multi-way sentiment analysis. In *ALTA-2012 – Proceedings of the Australasian Language Technology Workshop*, pages 52–60, Dunedin, New Zealand.

M. Craven and J. Kumlien. 1999. Constructing biological knowledge bases by extracting information from text sources. In *ISMB 1999 – Proceedings of the International Conference on Intelligent Systems in Molecular Biology*, pages 77–86, Heidelberg, Germany.

C. Jacob, P. Thomas, and U. Leser. 2013. Comprehensive benchmark of Gene Ontology concept recognition tools. In *Proceedings of BioLINK SIG 2013*, pages 20–26, Berlin, Germany.

A. Jimeno-Yepes, E. Jimenez-Ruiz, V. Lee, S. Gaudan, R. Berlanga, and D. Rebholz-Schuhmann. 2008. Assessment of disease named entity recognition on a corpus of annotated sentences. *BMC Bioinformatics*, 9(S-3).

J.-D. Kim, T. Ohta, Y. Tateisi, and J. Tsujii. 2003. Genia corpus – a semantically annotated corpus for biotextmining. *Bioinformatics*, 19(suppl. 1):i180–i182.

S. Kulick, A. Bies, M. Liberman, M. Mandel, R. McDonald, M. Palmer, A. Schein, L. Ungar, S. Winters, and P. White. 2004. Integrated annotation for biomedical information extraction. In *HLT/NAACL-2004 – Proceedings of the Human Language Technology Conference and the Annual Meeting of the North American Chapter of the Association for Computational Linguistics*, pages 61–68, Boston, Massachusetts.

J.D. Lafferty, A. McCallum, and F.C.N. Pereira. 2001. Conditional Random Fields: Probabilistic models for segmenting and labeling sequence data. In *ICML'2001 – Proceedings of the 18th International Conference on Machine Learning*, pages 282–289, Williamstown, Massachusetts.

R. Leaman and G. Gonzalez. 2008. BANNER: an executable survey of advances in biomedical named entity recognition. In R.B. Altman, A.K. Dunker, L. Hunter, T. Murray, and T.E. Klein, editors, *Pacific Symposium on Biocomputing*, pages 652–663. World Scientific.

R. Leaman, C. Miller, and G. Gonzalez. 2009. Enabling recognition of diseases in biomedical text with machine learning: corpus and benchmark. In *LBM2009 – Proceedings of the 3rd International Symposium on Languages in Biology and Medicine*, pages 82–89, Jeju Island, South Korea.

J.P. Pestian, C. Brew, P. Matykiewicz, D.J. Hovermale, N. Johnson, K. Bretonnel Cohen, and W. Duch. 2007. A shared task involving multi-label classification of clinical free text. In *BioNLP'07 – Proceedings of the Workshop on Biological, Translational, and Clinical Language Processing*, pages 97–104, Prague, Czech Republic.

S. Pyysalo, A. Airola, J. Heimonen, J. Björne, F. Ginter, and T. Salakoski. 2008. Comparative analysis of five protein-protein interaction corpora. *BMC Bioinformatics*, 9(Suppl. 3):S6.

D. Rebholz-Schuhmann, M. Arregui, S. Gaudan, H. Kirsch, and A. Jimeno-Yepes. 2008. Text processing through Web services: calling Whatizit. *Bioinformatics*, 24(2):296–298.

B. Rosario and M.A. Hearst. 2004. Classifying semantic relations in bioscience texts. In *ACL'2004 – Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics*, pages 431–438, Barcelona, Spain.

B. Settles. 2004. Biomedical named entity recognition using Conditional Random Fields and rich feature sets. In *Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and its Applications*, pages 104–107, Geneva, Switzerland.

P. Stenetorp, S. Pyysalo, G. Topic, T. Ohta, S. Ananiadou, and J Tsujii. 2012. BRAT: A Web-based tool for NLP-assisted text annotation. In *EACL'2012 – Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 102–107, Avignon, France.

K.M. Verspoor, K.B. Cohen, A. Lanfranchi, C. Warner, H.L. Johnson, C. Roeder, J.D. Choi, C. Funk, Y. Malenkiy, M. Eckert, N. Xue, W.A. Baumgartner, M. Bada, M. Palmer, and L.E. Hunter. 2012. A corpus of full-text journal articles is a robust evaluation tool for revealing differences in performance of biomedical natural language processing tools. *BMC Bioinformatics*, 13(107).

K.M. Verspoor, A. Jimeno-Yepes, L. Cavedon, T. McIntosh, A. Herten-Crabb, Z. Thomas, and J.P. Plazzer. 2013. Annotating the biomedical literature for the human variome. *Database: the journal of biological databases and curation*, 2013.