The impact of near domain transfer on biomedical named entity recognition

Nigel Collier*
European Bioinformatics Institute
Hinxton, Cambridge, UK, and
National Institute of Informatics, Tokyo, Japan

Ferdinand Paster
University of Applied Sciences
Upper Austria
Hagenberg Campus, Austria

Mai-vu Tran
University of Engineering and Technology - VNU
Hanoi, Vietnam

Abstract

Current research in fully supervised biomedical named entity recognition (bioNER) is often conducted in a setting of low sample sizes. Whilst experimental results show strong performance in-domain it has been recognised that quality suffers when models are applied to heterogeneous text collections. However the causal factors have until now been uncertain. In this paper we describe a controlled experiment into near domain bias for two Medline corpora on hereditary diseases. Five strategies are employed for mitigating the impact of near domain transference including simple transference, pooling, stacking, class re-labeling and feature augmentation. We measure their effect on f-score performance against an in domain baseline. Stacking and feature augmentation mitigate f-score loss but do not necessarily result in superior performance except for selected classes. Simple pooling of data across domains failed to exploit size effects for most classes. We conclude that we can expect lower performance and higher annotation costs if we do not adequately compensate for the distributional dissimilarities of domains during learning.

1 Introduction

Model and feature selection are important experimental tasks in supervised machine learning for suggesting approaches that will generalise well on real world data. Research in biomedical named entity recognition (bioNER) often displays two features: (1) small samples of labeled data, and (2) an implicit assumption that the future data will be drawn from a similar distribution to the labeled data and hence that minimising expected prediction error on held out data will minimise actual future loss. Since expert labeling is time consuming and expensive, labeled data sets tend to be relatively small, e.g. (Kim et al., 2003; Tanabe et al., 2005; Pyysalo et al., 2007), in the region of a few hundred or thousand Medline abstracts. Despite the danger of intrinsic idiosyncracies such corpora are often used to demonstrate putative prediction error across the heterogeneous collection of 22 million Medline abstracts. Once this assumption is made explicit it is of interest to both researchers and users that the implications and limitations of such experimental settings are explored.

Cross domain studies have indicated an advantage for mechanisms that compensate for domain bias. For fully supervised learning, which is the scenario we explore here, recent methods include: feature augmentation (Daumé III, 2007; Arnold et al., 2008; McClosky et al., 2010), instance weighting (Jiang and Zhai, 2007; Foster et al., 2010), schema harmonisation (Wang et al., 2010) and semi-supervised/lightly supervised approaches (Sagae and Tsujii, 2007; Liu et al., 2011; Pan et al., 2013). More generally there is a wide body of work in transfer learning (also known as *domain adaptation*) that tries to handle discrepancies between training and testing distributions (Pan and Yang, 2010).

As an illustration of near domain bias consider the list of high frequency named entities in Table 1 drawn from two sub-domains in the research literature of hereditary diseases. A domain expert in hereditary diseases would have no difficulty in dividing them into two non-overlapping sets corresponding to the two near domains with one term t_5 *patients* shared by both: $\{t_1,t_6,t_8,t_9\}$ and $\{t_2,t_3,t_4,t_7,t_{10}\}$.

Previous studies have shown what happens when you radically change the domain and/or the

*čollier@ebi.ac.uk

Proceedings of the 5th International Workshop on Health Text Mining and Information Analysis (Louhi) @ EACL 2014, pages 11–20,
Gothenburg, Sweden, April 26-30 2014.

t_1	rheumatoid arthritis	t_6	human leukocyte antigen
t_2	lupus erythematosus	t_7	coronary heart disease
t_3	leopard syndrome	t_8	type 1 diabetes
t_4	Omapatrilat	t_9	T1D
t_5	patients	t_{10}	hypertension

Table 1: High frequency entities in the hereditory disease literature for auto-immune and cardiovascular diseases.

annotation schema, e.g. from newswire to Medline or Web pages. But what happens when the annotation schema, the annotator and the primary domain stay the same? Although the notion of *domain* is difficult to formalise in the context of research literature, this study explores the condition where the variable factor is a shift to a *near domain* of literature as defined by biocurators and illustrated in the previous example. Our contribution to biomedical named entity recognition (bioNER) is in five areas:

1. We compare four data combination strategies for mitigating the impact of near domain transference and measure their effect on f-score performance against an in domain baseline.

2. We provide additional evidence for the effectiveness of (Daumé III, 2007)'s *frustratingly simple* strategy which provides both general and domain-specific features; in effect a joint learning model.

3. Expectedly, but not trivially, we show that a general loss of f-score occurs on bioNER when transfering to near domains. This loss is not uniform across all classes. We provide class-by-class drill down analysis to the underlying causal factors which make some entities more robust to near domain transference in biomedicine than others.

4. Our results challenge the notion that pooling small corpora, even when guideline differences are reconciled, leads to improved f-score performance (Wang et al., 2010; Wagholikar et al., 2013).

5. In addition to the usual biomedical entity types we introduce the class of phenotypes which are valued as indicators of genetic malfunction and characteristic of diseases. The phenotype class incorporates a complex dependency between classes, notably anatomical entities and genes.

This paper is organised as follows: Section 2 describes related work in cross domain transfer for biomedical NER, Section 3 discusses our approach including the two data sets used in our experiments, CRF model, feature choices and evaluation framework. In Section 4 we outline our experimental design. Finally in Section 5 we compare the performance of six data selection strategies that try to maximise f-score performance on domain entity classes in the target corpus.

2 Related work

It is surprising that there exists, to the best of our knowledge, no controlled study that has shed light on the issue of near domain transfer for bioNER in a straightforward manner. The closest approach to our investigation in the biomedical domain is (Wang et al., 2009). Wang et al. explore potential sources of incompatibility across major bioNER corpora with different annotation schema (GENIA - 2000 Medline abstracts, GENETAG - approximately 20,000 Medline sentences and AIMed - 225 Medline abstracts). They focus exclusively on protein name recognition and observe a drop in performance of 12% f-score when combining data from different corpora. Various reasons are put forwards such as differences in entity boundary conventions, the scope of the entity class definitions, distributional properties of the entity classes and the degree of overlap between corpora.

A follow up study by the authors (Wang et al., 2010) looked at increasing compatibility between the GENIA and GENETAG corpora by reorganising the annotation schema to unify protein, DNA and RNA NER under a new label GGP (Gene and Gene Product). However the best performance from the coarse grained annotations still do not improve on the intra-corpus data.

In earlier work, (Tsai et al., 2006) looked at schema differences between the JNLPBA corpus of 2000 Medline abstracts (Kim et al., 2004) and the BioCreative corpus of 15,000 Medline sentences (Yeh et al., 2005) and tried to harmonise matching criteria. They demonstrated that relaxing the boundary matching criteria was helpful in maximising the cross-domain performance.

In the clinical domain (Wagholikar et al., 2013), explore the effect of harmonising annotation guidelines on the 2010 i2b2 challenge with Mayo Clinic Rochester (MCR) electronic patient records. They concluded that the effectiveness of pooling - i.e. merging of corpora by ensuring a common format and harmonised semantics - is dependent on several factors including compatibility between the annotation schema and differences in size. Again they noticed that simple pooling resulted in a loss of f-score, 12% for MCR and 4% for i2b2. They concluded that the asymmetry was likely due to size effects of the corpora, i.e. MCR being smaller suffered a greater loss due to the classifier being biased towards i2b2.

Due to the formulation of these studies and their limited scope it has previously been difficult to understand the precise causual factors affecting performance. Our study sheds light on the expected level of loss under different combination strategies and more importantly highlights the non-uniform nature of that loss.

3 Approach

We assume two small labeled data sets $D^S = d_1^s..d_n^s$ and $D^T = d_1^t..d_m^t$. $d_i^s = \langle x_i \subset X, y_i \subset Y \rangle$ is drawn from an unknown distribution P^s and represents the source document examples. Similarly,$d_i^t = \langle x_i \subset X, y_i \subset Y \rangle$ is also drawn from an unknown distribution P^t and represents the target document examples. We assume that D^S has N examples and D^T has M examples where $N \approx M$. x_i represents a covariate or feature vector and y_i is a target or label that can take multiple discrete values. We have a learning algorithm that learns a function $h : X \rightarrow Y$ with minimal loss on the portion of D^T used for testing. Any combination of D^S and D^T which are not used in testing can be used to learn h. Our task is to explore various strategies for data selection and re-factoring labels/features in order to maximise held out performance.

3.1 Data

In this paper we aim to empirically test domain transferrence for bioNER under the condition that the test and training data are relatively small and drawn from near domains, i.e. from studies on different types of heritable diseases. To do this we selected Medline abstracts from PubMed that were cited by biocuration experts in the canonical database on heritable diseases, the Online Mendelian Inheritance of Man (OMIM) (Hamosh et al., 2005). We selected *auto-immune diseases* and *cardio-vascular diseases* for our two corpora which we denote as C1 and C2 respectively. By comparing performance of a single model, a single annotator and a single annotation scheme with a range of sampling techniques we hope to quantify the effects of domain transferrence in isolation.

The target classes for the entities are as follows:

ANA Anatomical structures in the body. e.g. *liver*, *heart*.

CHE A chemical or drug. e.g. *pristane*, *histamine*, *S-nitrosoglutathione*.

DIS Diseases. e.g. *end stage renal disease*, *mitral valve prolapse*.

GGP Genes and gene products. e.g. *KLKB1 gene*, *highly penetrant recessive major gene*.

PHE Phenotype entities describing observable and measurable characteristic of an organism. e.g. *cardiovascular abnormalities*, *abundant ragged-red fibers*, *elevated IgE levels*.

ORG A living organism. e.g.*first-degree relatives*, *mice*.

The two corpora were annotated by a single experienced annotator who had participated in the GENIA entity and event corpus annotation. We developed detailed guidelines for single span none-nested entities before conducting a training and feedback session. Feedback was conducted over two weeks by email and direct meetings with the annotator and then annotation took approximately two months. The characteristics of the two corpora are shown in Table 2. Because annotation was carried out by only one person we do not provide inter-annotator scores.

Importantly, we note four points at this stage: (1) We incorporate a new named entity type, phenotype, which is aligned with investigations into heritable diseases. Semantically it is interesting because phenotypes annotated in the auto-immune literature pertain more often to sub-cellular processes and those in the cardiovascular domain pertain more often to cells, tissues and organs; (2) It can be seen that two NE classes fall well below 500 instances - what we might arbitarily consider the necessary level of support for high levels of performance. These are ANA and CHE;

	C1	C2	a	b
Abstracts	110	80	-	-
Tokens	27,421	26,578	-	-
Av. length	32.57	29.93	-	-
ANA	194	195	0.33	0.26
	(138)	(133)		
CHE	44	147	0.08	0.07
	(33)	(75)		
DIS	892	955	0.39	0.27
	(282)	(442)		
GGP	1663	754	0.41	0.45
	(928)	(511)		
ORG	799	770	0.56	0.67
	(429)	(323)		
PHE	507	1430	0.52	0.33
	(423)	(1113)		

Table 2: Characteristics of the C1 auto-immune and C2 cardiovascular corpora: number of abstracts, number of tokens, average sentence length, frequency of each entity type. Figures in parentheses represent counts after removing duplication. a: probability that a word in an entity class X in C1 is also a word in entity class X in C2. b: probability that a word in an entity class X in C2 is also a word in entity class X in C1

(3) We calculated from Table 2 the average number of mentions for each entity form by class and noted that this is relatively stable across corpora, except for DIS which has less variation in C2 than C1 and CHE which has more variation in C2 than C1. When combining evidence from both corpora the approximate order of type/token ratio are $PHE < ANA < CHE, GGP < ORG < DIS$ indicating that on average PHE entities have the greatest variation. Average entity lengths in tokens (not shown) indicate that PHE are significantly longer than other entity mentions; and (4) We calculated the probability that a word token in an entity class from one corpus would appear in an instance of the same entity class in the other corpus, reported as columns a and b. Although the probability of an exact match in instances between entities in the two corpora is generally quite low (below 20% - data not shown) there appears to be significant vocabulary overlap in most classes except for chemicals.

3.2 Conditional Random Fields

As in (Finkel and Manning, 2009) we apply our approach to a linear chain conditional random field (CRF) model (Lafferty et al., 2001; McCallum and Wei, 2003; Settles, 2004; Doan et al., 2012) using the Mallet toolkit[1] with default parameters. CRFs have been shown consistently to be among the highest performing bioNER learners. The data selection strategies employed here though are neutral and could have been applied to any other fully supervised learner model.

3.3 Features

We made use of a wide range of features, both conventional features such as word or part of speech, as well as gazetteers derived from external classification schemes that have been hand crafted by experts. These are shown in Table 3. Previous studies such as (Ratinov and Roth, 2009) have noted that domain gazetteer features play a critical role in aiding classification. In order to show realistic model behaviour consistent with state-of-the-art techniques we have included gazetteers derived from: the Human Phenotype Ontology (HPO: 15,800 terms), the Mammalian Phenotype Ontology (MP: 23,700 terms), the Phenotypic Attribute and Trait Ontology (PATO: 2,200 synonyms), the Brenda Tissue Ontology (BTO: 9,600 synonyms), the Foundation Model of Anatomy (FMA: 120,000 terms), National Library of Medicine gene list (NLM: 9 million terms), UMLS disease terms (UMLS: 275,000 terms), Jochem chemical terms (JOCHEM: 320,000 terms).

The feature set is quite large and therefore there is a danger that the learner will be hindered. For feature selection, we conducted baseline test runs under the same experimental conditions as those reported here using a grid search on features F1 to F11 and found that f-score performance was uniformly lower when removing any feature (data not shown but available as supplementary material from the first author).

In order to characterise the contribution each feature is making in label prediction we wanted to provide a measure of similarity between the feature and the class label probability distributions. Here we use the Gain Ratio (GR) to estimate intra-corpus class prediction performance by each feature. GR was used as a splitting function in C4.5

[1] http://mallet.cs.umass.edu/

(Quinlan, 1993) and is defined as

$$GR(C, F) = IG(C, F)/H(F) \quad (1)$$

where C represents a class label and F represents a feature type. IG is information gain and defined as,

$$IG(C, F) = H(C) - H(C|X) \quad (2)$$

H is entropy and defined for feature types as,

$$H(F) = -\sum_{i=1}^{n} p(f_i) log_2(p(f_i)) \quad (3)$$

for n feature types $f_i \in F$. Further information can be found in (Quinlan, 1993). GR is used in C4.5 in preference to IG because of its ability to normalise for the biases in IG. Generally this results in GR having greater predictive accuracy than IR since it takes into account the number of feature values. Note that GR is undefined when the denominator is zero.

Several points emerge from looking at GR and IG values in Table 3:

- C1 (auto-immune) and C2 (cardio-vascular) have about the same information gain contribution from most features but C1 seems to benefit more from GENIA named entity tagging, Human Phenotype Ontology (HPO), Foundation Model of Anatomy (FMA) and Gene Ontology (GO) terms whereas C2 benefits more from the UMLS diseases and ChEBI terms.

- GO, containing terms about genetic processes, has a higher GR in C1 than C2. This supports what we already expected - that auto-immune diseases contain a higher proportion of information about genetic process phenotypes than cardiovascular.

- The GENIA POS tags seem to provide a slightly higher GR in C2 than in C1.

- Despite its large size, UMLS has a smaller GR on both corpora compared to some other resources like HPO or GO or MA. This is despite its high IG value.

3.4 Evaluation

Traditional re-sampling using *k-fold cross validation* (k-CV) divides the n labelled documents into k disjoint subsets of approximately equal size designated as D_i for $i = 1, .., k$. The NER learner is trained successively on $k - 1$ folds from D and tested on a held out fold over k iterations. In order to preserve independence between contexts in training and held out data we assume here that the unit of division is the document, i.e. a single Medline abstract. Estimated prediction error is calculated based on the learner's labels on the k held out folds. Whilst k-CV is known to be nearly unbiased it is a highly variable estimator. Several studies have looked at k-CV for small sample sets. For example, (Braga-Neto and Dougherty, 2004) found on classifier experiments for small microarray samples ($20 <= n <= 120$) that whilst k-CV showed low bias they suffered from excessive variance compared to bootstrap or resubstitution estimators.

One cause of variance has been identified as within-block and between-block training errors arising from the disproportionate effects of a single abstract appearing in the training set of many folds. In order to reduce this effect Monte Carlo cross validation was used (also called *CV with repetition*). 100 iterations were used to randomly reorder the documents in the corpora before 10-fold CV sampling was run (*cv10r100*). Sampling of documents is done without replacement so that the independence between training and testing sets are maintained. Stratification was not applied. Micro averaged f-scores for labeling accuracy were calculated based on the 1000 test folds for each model. Evaluation was done in both directions (training and testing) for each corpus C1 and C2 to show any asymmetrical effects. To minimse the time taken for each experiment a cluster computer was used with 48 nodes.

The matching criteria we employ is the exact match - i.e. the span of the system labeling and the held out data labels should be exactly the same. Although this is not a necessary criteria for some applications such as database curation we used it here as it is widely applied in shared evaluations and shows the clearest effects of modeling choice.

We evaluate using the named entity precision, recall and F-score calculated using the CoNLL 2003 Perl script. This was calculated as,

$$f - score = \frac{(2 \times precision \times recall)}{(precision + recall)} \quad (4)$$

where,

	Feature	$IG(C1, F_i)$	$GR(C1, F_i)$	$IG(C2, F_i)$	$GR(C2, F_i)$
F_1	Word	1.17	0.13	1.20	0.13
F_2	Lemma	1.15	0.13	1.18	0.13
F_3	POS tag	0.36	0.09	1.18	0.13
F_4	Chunk tag	0.22	0.12	0.26	0.10
F_5	GENIA NE[a]	0.20	0.35	0.14	0.27
F_6	Orthography	0.15	0.08	0.16	0.08
F_7	Domain prefix	0.11	0.11	0.11	0.10
F_8	Domain suffix	0.08	0.11	0.08	0.11
F_9	Word length	0.13	0.05	0.16	0.06
F_{10}	Parenthesis	0.04	0.20	0.04	0.23
F_{11}	Abbreviation	0.08	0.22	0.06	0.24
F_{12}	HPO[b]	0.07	0.41	0.09	0.33
F_{13}	MP[c]	0.03	0.33	0.06	0.33
F_{14}	PATO[d]	0.01	0.03	0.02	0.04
F_{15}	BTO[e]	0.03	0.32	0.03	0.29
F_{16}	FMA[f]	0.05	0.28	0.05	0.23
F_{17}	MA[g]	0.02	0.31	0.02	0.29
F_{18}	PRO[h]	0.02	0.12	0.03	0.15
F_{19}	ChEBI[i]	0.01	0.15	0.03	0.20
F_{20}	JOCHEM[j]	0.01	0.15	0.01	0.14
F_{21}	NCBI[k]	0.01	0.14	0.01	0.14
F_{22}	UMLS[l] disease	0.01	0.14	0.03	0.24
F_{23}	NCBI gene	0.02	0.18	0.02	0.19
F_{24}	GO[m]	0.13	0.38	0.05	0.28
F_{25}	UMLS[n]	0.48	0.12	0.52	0.11
F_{26}	45CLUSTERS[o]	0.50	0.10	0.47	0.10

Table 3: Features used in the experiments. [a]The GENIA named entity tagger (Kim et al., 2003), [b](Robinson et al., 2008), [c](Smith et al., 2004), [d](Gkoutos et al., 2005), [e](Gremse et al., 2011), [f](Rosse and Mejino, 2003), [g](Hayamizu et al., 2005), [h](Natale et al., 2011) , [i](Degtyarenko et al., 2008), [j](Hettne et al., 2009),[k](Federhen, 2012),[l](Lindberg et al., 1993),[m](Gene Ontology Consortium, 2000),[n]133 categories from the UMLS,[o]45 cluster classes derived by Richard Socher and Christoph Manning PubMed available at http://nlp.stanford.edu/software/bionlp2011-distsim-clusters-v1.tar.gz

$$precision = TP/(TP + FP) \quad (5)$$

and,

$$recall = TP/(TP + FN) \quad (6)$$

A true positive (TP) is a gold standard NE tagged by the system as an NE. A true negative (TN) is a gold standard none-NE tagged by the system as a none-NE. A false positive (FP) is a gold standard none-NE tagged by the system as an NE. Evaluation is based on correctly marked whole entities rather than tokens.

4 Experimental design

In this section we present the experimental conditions we used, starting with a description of the models which we designate M1 to M6 and describe below. All methods made use of 100 iterations of Monte Carlo 10-fold cross validation.

M1: IN DOMAIN We trained and tested on only the data for the source domain. This methods forms our baseline and represents the standard experimental setting.

M2: OUT DOMAIN We trained on the source domain and tested on the target domain. This method shows expected loss on near domain transferrence and represents the standard operational setting for users.

M3: MIX-IN We trained on 100% of the source domain and unified this with 90% of the folded in target domain data, leaving 10% for testing. This method reflects the pooling technique typically employed in corpus construction for bioNER.

M4: STACK We trained a CRF model on 100% of the source domain and stacked it with another CRF trained on 90% of the folded in target domain data. Stacking employs a meta-classifier and is a popular method for constructing high performance ensembles of classifiers (Ekbal and Saha, 2013). In this case we collected the output labels from the source domain-trained CRF on target sentences and added them as features for the target domain trained CRF.

M5: BINARY CLASS We re-labeled the complex class PHE as PHE-C1 in C1 and PHE-C2 in C2 and repeated M3. Afterwards we recombined PHE-C1 and PHE-C2 into PHE.

M6: FRUSTRATINGLY SIMPLE We followed the feature augmentation approach of (Daumé III, 2007). This method effectively provides a joint learning model on C1 and C2 by splitting each feature into three parts: one for sharing cross domain values and one for each domain specific value. We evaluated using the same regime as M3.

5 Experimental results and discussion

In Table 4 we show f-score performance from near biomedical domains with our six strategies. This section now tries to draw together an interpretation for the performance trends that we see and to drill down to some of the causal factors.

Held out tests performed in-domain (M1) on both corpora C1 and C2 indicate a relatively high level of performance, conservatively in line with state-of-the-art estimates. The broad trend in performance is for entity classes with more instances to out perform others with lower numbers. The class which most obviously breaks this trend is the complex entity type of PHE. To understand this consider that PHE is defined as an observable property on an organism and as such tends to be formed from a quality such as *malformed* that describes a structural entity such as *valve*. To see closer what is happening we looked at the confusion matrices for M1 on both corpora. For both C1 and C2 we observed that a substantial proportion of words inside PHE sequences were confused with GGP, DIS or ANA entities. Similarly a high proportion of words inside ANA sequences were confused with PHE entities. This indicates that dependencies within complex biomedical entities like PHE might better be modeled explicitly using tree-structures in a manner similar to events rather than using n-gram relations.

In the M2 out of domain experiments we see a generally severe loss of f-score performance across most classes. Training on C2 and testing on C1 results in a 19.1% loss (F1 69.9 to 50.8) and training on C1 and testing on C2 results in a 11.9% loss overall (F1 58.5 to 46.6). The results agree with Wang et al.'s experience on heterogeneous Medline corpora and extend the upper limit on all-class loss due to domain transferrence to 19%. The only NE class where we see a symmetric benefit from pooling entities in M3 is for ORG (F1 68.4 to 72.2, F1 73.2 to 77.4). Intriguingly the data from Tables 2 and 4 hint at a correlation between the success of M3 pooling for ORG and broad cross-domain compatibility on the vocabulary (over 50% of ORG vocabulary is shared across corpora). However this is not supported in the low sharing case for CHE where we see increased performance from pooling (F1 31.3 to 38.7) when the target is C2 but decreased performance when the target is C1 (F1 29.5 to 20.0).

When we look at the pooling method (M3) and compare to the in-domain method (M1) no obvious size effect occurs for the number of entities in each class. To see this we can examine entity classes with an imbalanced number of instances in C1 and C2 such as CHE, GGP and PHE. Consider the following three cases: (1) Adding 147 instances of CHE from C2 to 44 instances from C1 is associated with CHE performance dropping from M1:29.5 to M3:20.0 when tested on C1; (2) Similarly adding 1430 instances of PHE from C2 to 507 instances from C1 is associated with PHE performance dropping from 46.0 in M1 to 39.7 in M3 when tested on C1; (3) But adding 1663 instances of GGP from C1 to 754 from C2 is associated with GGP rising from 57.2 in M1 to 61.1 in M3. If simply pooling more entities was important to improved f-score we would expect to see a clearer pattern of improvement but we do not.

The overall pooling loss for all classes on M3 is within 3% in both directions and within the

Model	Target	ANA	CHE	DIS	GGP	PHE	ORG	ALL
M1	C1	**57.1**	29.5	**80.4**	74.0	**46.0**	68.4	**69.9**
M2	C1	34.3	26.9	57.7	55.6	26.9	64.0	50.8
M3	C1	50.8	20.0	77.9	71.7	39.7	*72.2*	67.3
M4	C1	56.3	17.4	79.0	**74.1**	44.1	*70.8*	69.8
M5	C1	56.7	**29.6**	77.3	72.7	41.5	**72.8**	68.3
M6	C1	**57.1**	27.7	79.0	73.4	44.9	*69.9*	69.5
M1	C2	37.2	31.3	72.9	57.2	46.5	73.2	58.5
M2	C2	21.2	20.2	57.0	52.3	24.4	68.5	46.6
M3	C2	36.8	*38.7*	72.3	**61.1**	44.0	*77.4*	**59.7**
M4	C2	34.8	*34.4*	72.5	*57.5*	45.9	74.7	58.5
M5	C2	34.1	**41.6**	**73.6**	*58.9*	43.2	**78.5**	**59.6**
M6	C2	**39.9**	*35.0*	*73.3*	56.4	**46.6**	*75.0*	*59.1*

Table 4: Named entity recognition f-scores using Methods 1 to 6. All methods were tested using 100 iterations of Monte Carlo 10-fold cross validation. Figures in bold show best in class scores. Figures in italics show scores above the M1 baseline.

bounds observed by (Wang et al., 2009) and (Wagholikar et al., 2013) for their pooling of heterogeneous Medline corpora. Except for the ORG class which we higlighted above, we might cautiously quantify the loss of pooled entity mentions as being in the range up to 9.5% for CHE but more typically below 4%. The majority of the differences they observed - which are not present in our data - are most likely due to concept definition differences and annotation conventions.

In contrast to our expectations the M4 experiments showed very mild benefits for stacking and these were mixed across entity types. M4 tests on C2 showed no general improvement but some improvement in CHE and ORG. M4 tests on C1 resulted again in no overall improvement except for some gain for ORG, supporting our hypothesis that there is greater compatibility in ORG across domains.

The M5 approach of splitting the PHE labels for the two corpora resulted in a noticable improvement over M3 on the C1 test but unfortunately this was not sustained when testing on C2.

It is striking that in the M6 experiments the feature augmentation method only just meets the in-domain f-score on C1 and mildly exceeds it on C2. One explanation is that the corpora are so small that a richer feature set has only marginal effects on performance. Table 3 certainly indicates that many of the features have low predictive capacity (gain ratio values below 0.1) in an intra-corpus setting but this is not the case for others such as GENIA NE tags or HPO gazzetteer terms.

Overall when we average the f-scores across models for C1 and C2 we see that there is a marginal benefit to the M1, M4 and M6 strategies over M3 and M5 with M2 suffering the greatest loss in performance.

6 Conclusion

In this paper we have provided evidence that transference even to closely related domains in biomedical NER incurs a severe loss in f-score. We have demonstrated empirically that strategies that make use of multi-domain corpora such as stacking learners and feature augmentation mitigate the accuracy loss but do not necessarily result in superior performance except for selected classes such as organisms where there appears to be broad terminology consensus. Simple pooling of data across domains failed to exploit size effects especially for the complex class of phenotypes. The list of strategies employed has not been exhaustive and it is possible that others such as feature hierarchies (Arnold et al., 2008) might yield better results.

BioNER is complicated by various factors such as descriptive names, polysemous terms, conjuctions, nested constructions and a high quantity of abbreviations. We have shown that performance is also held back by not considering document level properties related to domain such as topicality. We can expect lower performance and higher annotation costs if we do not adequately allow for the distributional dissimilarities of domains during learning, even in closely related topical settings.

Acknowledgments

The authors gratefully acknowledge the many helpful comments from the anonymous reviewers of this paper. Nigel Collier's research is supported by the European Commission through the Marie Curie International Incoming Fellowship (IIF) programme (Project: Phenominer, Ref: 301806).

References

A. Arnold, N. Nallapati, and W. Cohen. 2008. Exploiting feature hierarchy for transfer learning in named entity recognition. In *Annual meeting of the Association for Computational Linguistics (ACL 2008)*, pages 245–253.

U. Braga-Neto and E. Dougherty. 2004. Is cross-validation valid for small-sample microarray classification? *Bioinformatics*, 20(3):374–380.

H. Daumé III. 2007. Frustratingly easy domain adaptation. In *Annual meeting of the Association for Computational Linguistics (ACL 2007)*, pages 256–263.

K. Degtyarenko, P. de Matos, M. Ennis, J. Hastings, M. Zbinden, A. McNaught, R. Alcántara, M. Darsow, M. Guedj, and M. Ashburner. 2008. ChEBI: a database and ontology for chemical entities of biological interest. *Nucleic acids research*, 36(suppl 1):D344–D350.

S. Doan, N. Collier, H. Xu, P. Duy, and T. Phuong. 2012. Recognition of medication information from discharge summaries using ensembles of classifiers. *BMC Medical Informatics and Decision Making*, 12(1):36.

A. Ekbal and S. Saha. 2013. Stacked ensemble coupled with feature selection for biomedical entity extraction. *Knowledge-Based Systems*.

S. Federhen. 2012. The NCBI taxonomy database. *Nucleic acids research*, 40(D1):D136–D143.

J. Finkel and C. Manning. 2009. Hierarchical bayesian domain adaptation. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 602–610.

G. Foster, C. Goutte, and R. Kuhn. 2010. Discriminative instance weighting for domain adaptation in statistical machine translation. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing (EMNLP 2010)*, pages 451–459.

Gene Ontology Consortium. 2000. Gene ontology: tool for the unification of biology. *Nature Genetics*, 25:19–29.

G. Gkoutos, E. Green, A. Mallon, J. Hancock, and D. Davidson. 2005. Using ontologies to describe mouse phenotypes. *Genome Biology*, 6:R8.

M. Gremse, A. Chang, I. Schomburg, A. Grote, M. Scheer, C. Ebeling, and D. Schomburg. 2011. The BRENDA tissue ontology (BTO): the first all-integrating ontology of all organisms for enzyme sources. *Nucleic Acids Research*, 39(suppl 1):D507–D513.

A. Hamosh, A. F. Scott, J. S. Amberger, and C. A. Bocchini. 2005. Online mendelian inheritance of man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Research*, 33(suppl 1):D514–D517.

T. Hayamizu, M. Mangan, J. Corradi, J. Kadin, M. Ringwald, et al. 2005. The adult mouse anatomical dictionary: a tool for annotating and integrating data. *Genome Biol*, 6(3):R29.

K. Hettne, R. Stierum, M. Schuemie, P. Hendriksen, B. Schijvenaars, E. van Mulligen, J. Kleinjans, and J. Kors. 2009. A dictionary to identify small molecules and drugs in free text. *Bioinformatics*, 25(22):2983–2991.

J. Jiang and C. Zhai. 2007. Instance weighting for domain adaptation in NLP. In *Annual meeting of the Association for Computational Linguistics (ACL 2007)*, volume 2007, page 22.

J. D. Kim, T. Ohta, Y. Tateishi, and J. Tsujii. 2003. GENIA corpus - a semantically annotated corpus for bio-textmining. *Bioinformatics*, 19(Suppl.1):180–182.

J. Kim, T. Ohta, Y. Tsuruoka, Y. Tateisi, and N. Collier. 2004. Introduction to the bio-entity recognition task at JNLPBA. In N. Collier, P. Ruch, and A. Nazarenko, editors, *Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and its Applications (JNLPBA), Geneva, Switzerland*, pages 70–75, August 28–29. held in conjunction with COLING'2004.

J. Lafferty, A. McCallum, and F. Pereira. 2001. Conditional random fields: probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning, Massachusetts, USA*, pages 282–289, June 28th – July 1st.

Donald A.B. Lindberg, L. Humphreys, Betsy, and T. McCray, Alexa. 1993. The unified medical language system. *Methods of Information in Medicine*, 32:281–291.

Xiaohua Liu, Shaodian Zhang, Furu Wei, and Ming Zhou. 2011. Recognizing named entities in tweets. In *Annual meeting of the Association for Computational Linguistics (ACL 2011)*, pages 359–367.

A. McCallum and L. Wei. 2003. Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons. In *Proc. Seventh Conference on Natural language learning at HLT-NAACL 2003 - Volume 4*, CONLL '03, pages 188–191.

D. McClosky, E. Charniak, and M. Johnson. 2010. 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. Association for Computational Linguistics.

D. Natale, C. Arighi, W. Barker, J. Blake, C. Bult, M. Caudy, H. Drabkin, P. DEustachio, A. Evsikov, H. Huang, et al. 2011. The protein ontology: a structured representation of protein forms and complexes. *Nucleic acids research*, 39(suppl 1):D539–D545.

S. Pan and Q. Yang. 2010. A survey on transfer learning. *Knowledge and Data Engineering, IEEE Transactions on*, 22(10):1345–1359.

S. Pan, Z. Toh, and J. Su. 2013. Transfer joint embedding for cross-domain named entity recognition. *ACM Transactions on Information Systems (TOIS)*, 31(2):7.

S. Pyysalo, F. Ginter, J. Heimonen, J. Björne, J. Boberg, J. Järvinen, and T. Salakoski. 2007. Bioinfer: a corpus for information extraction in the biomedical domain. *BMC bioinformatics*, 8(1):50.

J. Quinlan. 1993. *C4. 5: programs for machine learning*, volume 1. Morgan kaufmann.

L. Ratinov and D. Roth. 2009. Design challenges and misconceptions in named entity recognition. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning (CoNLL)*, pages 147–155.

P. N. Robinson, S. Kohler, S. Bauer, D. Seelow, D. Horn, and S. Mundlos. 2008. The human phenotype ontology: a tool for annotating and analyzing human hereditary disease. *The American Journal of Human Genetics*, 83(5):610–615.

C. Rosse and J. L. V. Mejino. 2003. A reference ontology for bioinformatics: the Foundational Model of Anatomy. *Journal of Biomedical Informatics*, 36(6):478–500, December. PMID: 14759820.

K. Sagae and J. Tsujii. 2007. Dependency parsing and domain adaptation with lr models and parser ensembles. In *Conference on Empirical Methods in Natural Language Processing Conference on Computational Natural Language Learning (EMNLP-CoNLL)*, volume 2007, pages 1044–1050.

B. Settles. 2004. Biomedical named entity recognition using conditional random fields. In *Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and its Applications (JNLPBA) at COLING'2004, Geneva, Switzerland*, pages 104–107, August 28–29.

C. L. Smith, C. W. Goldsmith, and J. T. Eppig. 2004. The mammalian phenotype ontology as a tool for annotating, analyzing and comparing phenotypic information. *Genome Biology*, 6:R7.

L. Tanabe, N. Xie, L. H. Thom, W. Matten, and W. J. Wilbur. 2005. GENETAG: a tagged corpus for gene/protein named entity recognition. *BMC Bioinformatics*, 6(Suppl 1):S3.

R. Tsai, S. Wu, W. Chou, Y. Lin, D. He, J. Hsiang, T. Sung, and W. Hsu. 2006. Various criteria in the evaluation of biomedical named entity recognition. *BMC bioinformatics*, 7(1):92.

K. Wagholikar, M. Torii, S. Jonnalagadda, H. Liu, et al. 2013. Pooling annotated corpora for clinical concept extraction. *J. Biomedical Semantics*, 4:3.

Y. Wang, J. Kim, R. Sætre, S. Pyysalo, and J. Tsujii. 2009. Investigating heterogeneous protein annotations toward cross-corpora utilization. *BMC bioinformatics*, 10(1):403.

Y. Wang, J. Kim, R. Sætre, S Pyysalo, T. Ohta, and J. Tsujii. 2010. Improving the inter-corpora compatibility for protein annotations. *Journal of bioinformatics and computational biology*, 8(05):901–916.

A. Yeh, A. Morgan, M. Colosimo, and L. Hirschman. 2005. Biocreative task 1a: gene mention finding evaluation. *BMC bioinformatics*, 6(Suppl 1):S2.