Semi-supervised Prediction of Protein Interaction Sentences Exploiting Semantically Encoded Metrics

Tamara Polajnar and Mark Girolami

University of Glasgow, Glasgow, Scotland, G12 8QQ
tamara@dcs.gla.ac.uk,
WWW home page: http://www.dcs.gla.ac.uk/inference/

Abstract. Protein-protein interaction (PPI) identification is an integral component of many biomedical research and database curation tools. Automation of this task through classification is one of the key goals of text mining (TM). However, labelled PPI corpora required to train classifiers are generally small. In order to overcome this sparsity in the training data, we propose a novel method of integrating corpora that do not contain relevance judgements. Our approach uses a semantic language model to gather word similarity from a large unlabelled corpus. This additional information is integrated into the sentence classification process using kernel transformations and has a re-weighting effect on the training features that leads to an 8% improvement in F-score over the baseline results. Furthermore, we discover that some words which are generally considered indicative of interactions are actually neutralised by this process.

1 Introduction

Lack of fully annotated training data is one of the major bottlenecks in biomedical text mining. Even for PPI detection, which is one of the most investigated TM problems, there are only a few standard data sets. The usefulness of these data sets is limited by their size and annotation schema [6, 3, 22]. In this paper we present a new method that integrates unlabelled data in order to improve performance of a classifier trained on a smaller minimally annotated data set.

A PPI is a relation between two protein entities linked by an action descriptor, which is usually either a verb or a present (*-ing*) or past (*-ed*) participial adjective. Identification of interactions requires significant biological knowledge. In addition, annotation may also require grammatical expertise, depending on whether entities, interaction identifiers, or even sentence parse trees are considered. Therefore, the simplest kind of annotation possible is the one where the segments of texts are simply marked for relevance by the biologists. This type of labelling is useful for training algorithms that detect passages that contain PPIs as a first step in a full interaction extraction pipeline [14]. We use the AImed data set in which the protein entities are annotated and interacting pairs are specified [3]. We use the pairs annotation to judge which sentences contain interactions. The AImed corpus is emerging as standard and is being used in a variety of ways [8, 1], yet it only contains less than 2000 sentences.

Attempts to overcome this shortage in labelled data usually involve semi-supervised learning where samples without class labels are added to the training set [8]. This approach generally leads to greatest improvements in classification performance when there are few labelled sentences and many unlabelled sentences. However, semi-supervised learning is also volatile, and could lead to a significant loss in accuracy [23]. Furthermore, the underlying assumption is that the labelled and unlabelled data come from the same distribution. Unfortunately, this prevents us expanding a fully labelled corpus by combining corpora created by other queries.

In order to address these concerns, we present a novel method of integrating unlabelled data into the classification process. We first create a word-word co-occurrence matrix from a large unlabelled corpus through unsupervised means. This corpus has a related topic and contains the words from the training set vocabulary. The matrix is then used to re-weight the words in the sentence documents according to their meaning in the larger corpus, thereby including external information into the training process implicitly.

We consider two semantic representations, the Hyperspace Analogue to Language (HAL) [17, 5, 4] and the Bound Encoding of the Aggregate Language Environment (BEAGLE) [11, 12]. Both HAL and BEAGLE model semantic memory using co-occurrence of words within a defined context window. Therefore they are slightly different from Latent Semantic Analysis (LSA) [15] which is based in the word-document space.

Statistical word co-occurrence information has been successfully used for synonym identification and word-sense disambiguation [20], as well as query expansion in information retrieval [24, 2]. We are not aware of any previous work that uses these semantic models to integrate external knowledge into the classification process. However, the Wikipedia[1] corpus has been previously used, with LSA, to improve the semantic linking of words to aid in classification of news texts. The results did not show any improvement over linear classification methods [19].

In this paper, we show, for the first time, that this type of knowledge can help enhance classification in the document-document space used by the kernel classifiers. We gain statistically significant improvements in classification by incorporating the semantic matrices into the kernel space. In addition, we obtain significant insights into the word usage and importance of particular features in classification. These initial experiments show that interesting results can be achieved through exploitation of the complexity of biomedical terms. Semantic models, such as HAL and BEAGLE, can help explore linguistic phenomena like polysemy, that in general make biomedical text mining more difficult than text processing in other domains [14].

[1] http://wikipedia.org

2 Semantic Spaces

Semantic spaces were initially introduced as a way of modelling psycholinguistic phenomena such as language acquisition and semantic priming. More recently semantic models have been applied to and tailored for natural language processing tasks, resulting in a proliferation of models [20]. We use the semantic models to improve kernel-based classification techniques. We do this by constructing word similarity matrices based on HAL and BEAGLE and then incorporating them into the kernels as described in Sect. 3.3.

Both HAL and BEAGLE calculate the co-occurrence between a **target word**, t, and the words within a specified **context**. The context can be defined as a document, a sentence, a window of words, or even a path in a dependency parse tree, anchored at the target [20]. In HAL it is defined as a sliding window where the target is the last word in the window, while in BEAGLE it is the sentence containing the target word.

The words within the context are called the **basis**, b. The set of all target words, T, and the set of all basis words, B, are not necessarily equivalent. In general, the co-occurrence models are created by counting the number of times a basis occurs in the context of a target word. These counts are recorded in a $|T| \times |B|$ matrix, where the targets are represented by the row vectors, while the basis correspond to the columns.

Semantic models also include a vector space distance metric that is used to calculate the similarity between target row vectors. In classification, the data are encoded as vectors of features, representing points in some multi-dimensional space. The kernel, $k(\mathbf{x}_i, \mathbf{x}_j) = \phi(\mathbf{x}_i)^T \phi(\mathbf{x}_j)$, is a function that takes these data vectors and transforms them into a linear product space which represents the distances between the points. We investigate the use of two kernel functions, commonly employed for text classification, to calculate the distance between the word vectors. The cosine kernel is defined as $k_c(\mathbf{x}_i, \mathbf{x}_j) = \frac{\mathbf{x}_i \cdot \mathbf{x}_j}{|\mathbf{x}_i||\mathbf{x}_j|}$, and the Radial Basis Function (RBF) as $k_r(\mathbf{x}_i, \mathbf{x}_j) = exp(-\theta|\mathbf{x}_i - \mathbf{x}_j|^2)$.

2.1 Hyperspace Analogue to Language

The HAL matrix, $\mathbf{H}$, is constructed by passing a window of fixed length, L, across the corpus. The last word in the window is considered the target and the preceding words are the basis. Because the window slides across the corpus uniformly, the basis words are previous targets, and therefore $T = B$.

The strength of the co-occurrence between a target and the basis depends on the distance between the two words, l, $1 < l < L$, within the window. The co-occurrence scoring formula, $L - l + 1$, assigns lower significance to words that are further apart. The overall co-occurrence of a target-basis pair is the sum of the scores assigned every time they coincide within the sliding window, across the whole corpus.

Even though the matrix is square, it is not symmetric. In fact the transpose of the matrix reflects the co-occurrence scores between the target and the basis that occur within the window of length L *after* the target. Thus $\mathbf{H}$ and $\mathbf{H}^T$ together reflect the full context surrounding a target. There are two ways of combining this information so that it would be considered when the distance between targets is calculated. The first way is to concatenate $\mathbf{H}$ and $\mathbf{H}^T$ to produce a $|T| \times 2|B|$ matrix. The second way is to add the two matrices together $\mathbf{H}$+$\mathbf{H}^T$. We found that for our kernel combination method that the latter strategy is more effective. This was also the case when HAL was employed for query expansion [24], Therefore, from now on when we refer to $\mathbf{H}$ we will assume $\mathbf{H} = \mathbf{H} + \mathbf{H}^T$.

2.2 Bound Encoding of the Aggregate Language Environment

The BEAGLE model [11, 12] was proposed as a combined semantic space that incorporates word co-occurrence and word order. For the purpose of comparison with HAL, we only consider the word co-occurrence construction.

BEAGLE differs from HAL in that it does not use the raw word counts directly. Instead, it represents each target t with a $1 \times D$ signal vector, $\mathbf{e}(t)$, of points drawn from the Gaussian distribution $\mathcal{N}(0, (\frac{1}{\sqrt{D}})^2)$. The number of dimensions D is chosen manually so that it is large enough to ensure that this vector is unique for each target or basis word, yet small enough to reduce the burden on memory. It is suggested in [11] that multiples of 1024 are an appropriate choice for D, and they use $D = 2048$ to encode larger corpora. D is generally much smaller than the number of basis words in a large corpus, so this representation also provides a more compact encoding.

The context in BEAGLE is made of the basis words that occur in the same sentence as the target word. The target vectors in the BEAGLE co-occurrence matrix, $\mathbf{B}$, are sums of the environmental vectors of the basis words that occur within the context of the target word. The more times that a certain basis is found in the same sentence as the target, the stronger its signal will be within the vector $\mathbf{B}[t]$.

BEAGLE Cosine	**BEAGLE RBF**	**HAL Cosine**	**HAL RBF**
tnf	tnf	tnf	tnf
capacities	treated	glutamic	slightly
architectu	cip	egg	fra
biofunctio	angiotensi	slightly	vector
shptp	testament	fra	progressio
myogenic	subjected	bind	hearts
increases	activated	uninfected	augmented
inhibitors	immunodefi	vector	indirectly
bcl	mol	progressio	searched
immobilize	transfecti	hearts	diagnosis

Table 1. Examples of the top ranked words similar to *TNF (tumor necrosis factor)*. Definition of TNF from RefSeq: *This cytokine is involved in the regulation of a wide spectrum of biological processes including cell proliferation, differentiation, apoptosis, lipid metabolism, and coagulation. This cytokine has been implicated in a variety of diseases, including autoimmune diseases, insulin resistance, and cancer.*

3 Methods

We assess the performance of the semantic kernels using the Gaussian process (GP) classifier [9]. We have previously found that GPs outperform the support vector machine [10] on the AImed [3] data set for the task of PPI sentence detection [21].

We formulate the interaction detection problem as a PPI sentence classification task. This allows us to use bag-of-words (BOW) [16] features with which we can examine the information gain from semantic kernels. In addition, the baseline features we employ are easier to extract and require no annotation. We also use protein names as features. While we rely on gold standard annotations, the proteins could be also automatically annotated.

3.1 Corpora

We use the AImed [3] data set for classifier training and testing and the GENIA [13] corpus to construct the semantic models.

AImed has been used in multiple studies recently for exact interacting pair extraction [3, 8, 1]. It is rapidly becoming one of the standard data sets for PPI classification.

AImed has nearly 55,000 words and is annotated with PPIs. On the other hand, the larger GENIA corpus has over 432,000 words, which was constructed from MEDLINE queries: *human, blood cell*, and *transcription factor*. It is only annotated with named entities including proteins, thus the information in GENIA cannot be directly used for PPI classification. Consequently, any relevant subset of MEDLINE would be equally as useful for this task. The protein names can be found automatically and therefore the annotations in GENIA are not strictly necessary.

3.2 Features

We consider two types of features for this task, *short* and *protein*.

In *short*, each feature corresponds to a word. The words are defined as a sequence of letters limited to the length of ten characters as in [7] . We also used full words, including any that contained numbers and letters. Unfortunately, this technique led to lower classification performance, and therefore we do not report detailed results here.

For *protein* features, the basic word extraction technique is the same as for *short*. However, we substitute the manually annotated protein names in the AImed corpus with place holder strings enumerating each of the proteins in the sentence. Thus, in each sentence the first protein is named *ptngne*1, the second is *ptngne*2 and so on. This method effectively anonymises the proteins across the whole corpus, turning the sentences into patterns.

3.3 Kernel Construction

The target words used for the construction of the semantic matrices are the words occurring in the AImed data set. For BEAGLE the basis are all words that occur in the sentences with the target words, while in HAL the basis are the same as the target words. Some features that occur in AImed cannot be found in GENIA. During the construction of the HAL matrix we find some empty rows, which can cause problems during similarity calculations. We add a small scalar value to the entire matrix to avoid this problem.

The baseline classification results were obtained with the k_c and k_r (as defined in Sect. 2) kernels directly on the sentence data from the AImed corpus, $\mathbf{X} = \mathbf{x}_1, \ldots, \mathbf{x}_M$. M is the number of sentences in $\mathbf{X}$ and N is the number of features, *i.e.* the length of the vectors $\mathbf{x}$. The $N \times N$ HAL and BEAGLE word-similarity matrices were constructed using the semantic co-occurrence matrices generated from the GENIA corpus and transformed by the kernel functions, for example $\mathbf{H}_c = \{k_c(\mathbf{h}_i, \mathbf{h}_j)\}_{i,j=1}^{N}$. The sentence-sentence kernels are then constructed so that they include the word similarity matrix, for example $\mathbf{K}_{ij} = \mathbf{x}_i \mathbf{H}_c \mathbf{x}_j$ is the $HAL + cosine$ kernel for sentence classification.

3.4 Experiment Description

In order to effectively use HAL and BEAGLE as kernels, we need to determine initial settings for the comparison experiment. We examined the effects of different distance metrics, parameters, and window sizes ($L = 1 \ldots 30$) for HAL for several feature types on the AImed corpus. We investigated the effects that the number of dimensions, D, and the cosine and RBF distance metrics have on BEAGLE. In [11] claim that if it is large enough, *i.e.* $D > 1000$, the lists of similar words produced do not change. Nevertheless, similarity values will make a difference in our experiments, so it is a parameter worth considering. We tested for $D = \{2048, 4096\}$. In Sect. 4 we report the observations gathered from these intial experiments and then present further experiments using the best results for each of the methods. The initial experiments for HAL encompassed a wide search space and as such were only ten-fold cross-validations. On the other hand, since the search space was much smaller, the final comparison results are an average of ten ten-fold cross-validations.

3.5 Evaluation Measures

Results were evaluated using the error (E), precision (P), recall (R), and F measures, which are defined in terms of true positives (tp), false positives (fp), true negatives (tn), and false negatives (fn) as follows: $E = \frac{fp+fn}{tp+tn+fp+fn}$, $P = \frac{tp}{tp+fp}$, $R = \frac{tp}{tp+fn}$, $F = \frac{2 \cdot P \cdot R}{P+R}$ [25]. The area under the receiver operator characteristic (ROC) curve is also employed as a standard measure. The ROC is a plot of the true positive rate vs. the false positive rate, and the larger the area under the curve (AUC) the better the performance of the classifier. When perfect classifier performance is achieved the AUC is 1. We also provide the average of the predictive likelihood (PL) for each of the cross validation experiments.

4 Experimental Results

Cosine Kernel

settings	results
features:	†F = 0.5384 ± 0.0049
short	E = 23.1394 ± 0.2890
kernel:	P = 0.7186 ± 0.0065
cosine	R = 0.4346 ± 0.0060
	†AUC = 0.7934 ± 0.0034
	PL = 0.0315 ± 0.0036

settings	results
features:	†F = 0.6789 ± 0.0043
protein	E = 18.6717 ± 0.2460
kernel:	P = 0.7258 ± 0.0056
cosine	R = 0.6414 ± 0.0057
	†AUC = 0.8688 ± 0.0025
	PL = 0.1341 ± 0.0038

HAL Kernel

settings	results
features:	†F = 0.5750 ± 0.0055
short	E = 23.6515 ± 0.2850
L:	P = 0.6482 ± 0.0068
8	R = 0.5197 ± 0.0060
kernel:	†AUC = 0.7820 ± 0.0034
H + RBF	PL = 0.0241 ± 0.0047

settings	results
features:	**†F = 0.7267 ± 0.0040**
protein	E = 16.3737 ± 0.2296
L:	P = 0.7514 ± 0.0055
1	R = 0.7061 ± 0.0048
kernel:	†AUC = 0.8953 ± 0.0022
H + RBF	PL = 0.2237 ± 0.0055

BEAGLE Kernel

settings	results
features:	**†F = 0.6167 ± 0.0052**
short	E = 21.6869 ± 0.2566
D:	P = 0.6801 ± 0.0064
2048	R = 0.5671 ± 0.0059
kernel:	†AUC = 0.7997 ± 0.0033
B + RBF	PL = 0.0555 ± 0.0049

settings	results
features:	†F = 0.7103 ± 0.0043
protein	E = 17.3131 ± 0.2535
D:	P = 0.7378 ± 0.0061
4096	R = 0.6880 ± 0.0051
kernel:	†AUC = 0.8895 ± 0.0022
B + cosine	PL = 0.2110 ± 0.0055

Table 2. Average results over ten ten-fold cross-validation experiments where the best settings for each of the methods were used. Two types of features were examined, plain words concatenated to the maximum of ten letters (*short*) and the same feature set but with protein names replaced by place holder strings (*protein*). The † indicates that all F-scores and AUCs are significantly different from all the other results using the same features.

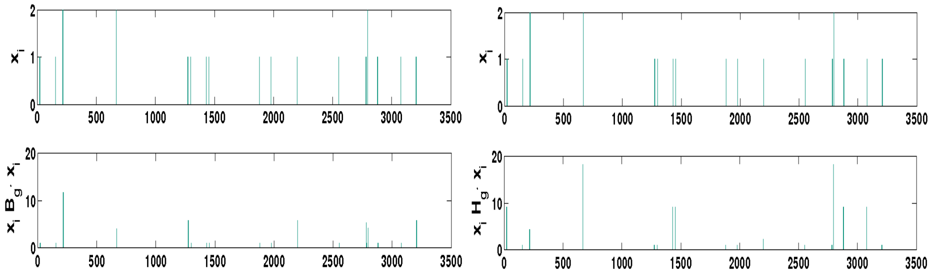
4.1 Experimental Parameter Selection

We found that for the sentence classification without semantic information, the cosine kernel always gave a higher F-score than the RBF. Therefore, we use the results obtained using the cosine kernel as the baseline for comparison with the semantic kernels.

The experiments to find the right parameters for the HAL kernel were conducted in two stages. Firstly, we found that the scalar value that is added to the matrix **H**, to prevent division by zero while performing similarity transformations, does not have any influence on classification. In addition, testing shows that RBF parameter θ makes little difference when the kernel is applied to the HAL and BEAGLE matrices.

Next, we tested which of the similarity measures will give the highest classification results, for each of the window sizes. We found that the contents of the HAL matrix are highly influenced by the choice of window width parameter L. The right choice of L and the similarity metric could give variations of over 5% in the F-score. We chose three sets of parameters for further experimentation: the ones that gave the highest F-score, the highest AUC, and the lowest error.

Unlike HAL, the co-occurrence component of BEAGLE has only one parameter D, resulting in a smaller search space. In general, we found that for



`Word indices: (21) tnfr, (153) tnf, (216) ligand, (667) human, (1274) discovered, (1298) designated, (1430) protein, (1453) glucocorti, (1879) ortholog, (1977) induced, (2199) recently, (2551) hgitrl, (2780) identified, (2785) receptor, (2797) related, (2881) hgitr, (3079) gitr, (3207) murine.`

Fig. 1. Re-weighting of words in a sentence by the BEAGLE and HAL kernels. This figure demonstrates the neutralisation of some features while others are given higher importance.

BEAGLE the length D of the signal vector $e(t)$ has a lesser effect than the choice of similarity metric.

4.2 The Effects of HAL and BEAGLE on Target Words

The word-similarity lists that semantic spaces produce are difficult to evaluate quantitatively. For biomedical texts, there are no large-scale user-driven linguistic study results that could be used to evaluate these types of lists. For example, Table 1 shows lists of the most similar words to *TNF* from both the HAL and BEAGLE matrices as transformed by the two similarity metrics. It is obvious that there are differences in the lists, however it is difficult to quantify which list is the best. TNF is a cytokine that is involved in several essential cellular processes and consequently it appears to be a key factor in many diseases including cancer. There are many studies that evaluate TNF interactions and their consequences. The different similarity lists appear to reflect some of the types of different articles written. For example, the BEAGLE matrix transformed by cosine, $\mathbf{B}_c$, tends to weight highly the words that have to do with the function of TNF in different organs. This is supported by the fact that, the words *liver* and *kidney* appear further down the list, at positions 11 and 18, respectively. The lists produced by the BEAGLE with RBF ($\mathbf{B}_r$) and HAL with cosine ($\mathbf{H}_c$) similarity matrices reflect more of a biomolecular experimental view, while the list from $\mathbf{H}_r$ appears to contain more words that would be found in clinical medical abstracts.

4.3 The Effects of HAL and BEAGLE on Sentences

When we examine the similarity vectors of individual words within the HAL and BEAGLE spaces we find that some words are highly similar to many other targets while others are only similar to themselves. Due to the way that each of the sentences is multiplied by the similarity vector, the sum of the similarity values for each of the target words becomes the key. For example, if we concentrate on the similarity space created from GENIA, using *short* features and the RBF similarity metric, we can observe the transformations that happen to a single sentence from the AImed corpus. So, from the sentence:

```
    We have identified a new TNF - related ligand , designated human GITR ligand ( hGITRL ) ,
and its human receptor ( hGITR ) , an ortholog of the recently discovered murine glucocorticoid
- induce d TNFR - related ( mGITR ) protein [ 4 ] .
```

we can extract the following vector $\mathbf{x}_1$ represented by non-zero fetures: *tnfr:1, tnf:1, discovered:1, designated:1, protein:1, glucocorti:1, otholog:1, induced:1, recently:1, hgitrl:1, identified:1, receptor:1, hgitr:1, gitr:1, murine:1, ligand:2, human:2, related:2.*

In general, it would be highly correlated with other sentences that contain these same words in high proportions.

However, after including the global knowledge encoded in the $\mathbf{B}_r$ kernel, we found that these values were greatly altered. If the sentence contains features that are related to many others the similarity with itself will be higher, but also these words will be boosted in significance when calculating the inner product with other sentence vectors. So for $\mathbf{x}_1$, after transformation we got $\mathbf{x}_1\mathbf{B}_r\mathbf{x}_1^T = 53.7142$. The features in the sentence were weighted as follows: *designated:1, receptor:1, hgitrl:1, protein:1, induced:1, gitr:1, ortholog:1, tnfr:1, tnf:1.0055, glucocorti:1.0492, hgitr:1.0533, human:4.0001, related:4.1569, identified:5.3208, murine:5.8166, discovered:5.8180, recently:5.8195, ligand:11.6744.* We can visualise this transformation in Fig. 1 for both the BEAGLE and HAL kernels. This is an example of an entry on the diagonal of the kernel, but the same calculations were made between any two sentences, *e.g.* $\mathbf{x}_1\mathbf{B}_r\mathbf{x}_3^T = 23.3594$.

4.4 The Effects of BEAGLE and HAL on Classification

Incorporation of semantic information from the HAL and BEAGLE matrices significantly increases the classification performance (Table 2). With the basic *short* features we find that the BEAGLE matrix with RBF similarity increases the F-score by nearly 8%. When employing *protein* features we see less of an improvement, though it is still statistically significant. Using HAL with RBF similarity leads to 5% improvement in the F-score.

4.5 Feature Re-weighting and Classification Performance

In order to understand the increase in performance we have to examine the effects of the kernels on the features. In general, the RBF kernel produces a sparser kernel with higher contrast, *i.e.* sharper decline in similarity values. This can also be observed by examining the highest weighted word in the $\mathbf{B}_r$ matrix, *asp* and one of the lowest weighted, *protein.* Their weight vectors are plotted in Fig. 2.

Protein is one of the words that is generally considered to be an indicator of interactions For example, [18] use a list of 83 discriminating words to score abstracts according to their presence or absence. Some of the top words they use are: *complex, interaction, two-hybrid, interact, proteins, protein, domain, interactions, required, kinase, interacts, complexes, function, essential, binding, component, etc.*

We find that $\mathbf{B}_r$ kernel actually reduces the weight for many of these words. For example, *complex, interaction, interact, protein, binding, domain, kinase, complexes,* and *function* all get multiplied only by factor of 1. This implies that these words are only similar to themselves. However, other words including *hybrid, proteins, required, interacts, essential,* and *component* get multiplied by numbers orders of magnitude larger, for example 800, implying high similarity with many words. This has the effect of drastically reordering the significance of words in a way that cascades into the final sentence-sentence similarity space.

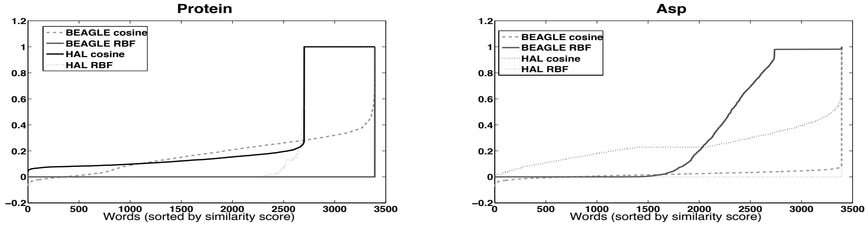


Fig. 2. Similarity calculations between the chosen words and the rest of the lexicon as calculated by the different kernels. This figure demonstrates the neutralising effect of the BEAGLE kernel on the high-frequency word *protein*.

When we examine the properties of the AImed corpus we can see the advantages of the $\mathbf{B}_r$ scaling. The most frequent words in the positive data are: *binding, protein, receptor, interactio, il, beta, domain, complex, cells, human, cell, kinase, ...*, while the top negative words are: *protein, receptor, cell, binding, cells, human, proteins, il, transcript, interactio, domain, expression,...* Therefore we can gather that, actually, for this data there is a large intersection of positive and negative high-frequency words, and thus they are not very discriminative. On the other hand, the words that occur more in the positive data than in negative are: *interacts, binds, complex, hsp, gp, ccr, cdk, ...*; so, the higher weights assigned to these words improve classification.

5 Discussion

In this paper, we have presented a new method of integrating unlabelled data, via kernel substitution, in order to improve supervised classification performance. We use the unsupervised semantic models to combine word usage information from a large external corpus into the kernel space. With this method we are able to integrate data that does not necessarily come from the same distribution as the training data, which is a requirement of traditional semi-supervised approaches.

Integration of word co-occurrence data in this manner leads to almost an 8% improvement in the F-score on BOW features and a 5% improvement when using protein annotations in the feature set.

This is the first time HAL and BEAGLE semantic spaces have been combined within a kernel classifier in this way. These models re-introduce the semantic links that had been originally lost through the choice of BOW features. By re-weighting the words in a sentence, these models emphasise terms that have many synonyms and thus are more interchangeable with terms that occur in other sentences. Therefore by equating semantically synonymous terms we were able to increase classification performance. The same type of improvement was observed when we artificially anonymised the proteins by substituting a placeholder string for a protein name. However, the proposed semantic models are unsupervised and not limited to handling only manually chosen entity types.

These initial experiments introduce new avenues of research that can be undertaken to further explore unlabelled data integration through the kernel space.

6 Acknowledgements

TP is funded by a Scottish Enterprise PhD studentship. MG is funded by an EPSRC Advanced Research Fellowship EP/E052029/1 and EPSRC project CLIMB EP/F009429/1.

References

1. A. Airola, S. Pyysalo, J. Björne, T. Pahikkala, F. Ginter, and T. Salakoski. All-paths graph kernel for protein-protein interaction extraction with evaluation of cross-corpus learning. *BMC bioinformatics*, 9 Suppl 11, 2008.
2. Leif Azzopardi, Mark Girolami, and Malcolm Crowe. Probabilistic hyperspace analogue to language. In *SIGIR '05: Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 575–576, New York, NY, USA, 2005. ACM.
3. R Bunescu, R Ge, R J Kate, E M Marcotte, R J Mooney, A K Ramani, and Y W Wong. Comparative experiments on learning information extractors for proteins and their interactions. *Artif Intell Med*, 33(2):139–155, Feb 2005.
4. C Burgess, K. Livesay, and K. Lund. Explorations in context space: Words, sentences, discourse. In *Discourse Processes*, volume 25, pages 211 – 257. 1998.
5. C. Burgess and K Lund. Modeling parsing constraints with high-dimensional context space. In *Language and Cognitive Processes*, volume 12, pages 177–210. 1997.
6. K. Bretonnel Cohen, Lynne Fox, Philip V. Ogren, and Lawrence Hunter. Corpus design for biomedical natural language processing. In *Proceedings of the ACL-ISMB workshop on linking biological literature, ontologies and databases: mining biological semantics*, pages 38–45, 2005.
7. Ian Donaldson, Joel Martin, Berry de Bruijn, Cheryl Wolting, Vicki Lay, Brigitte Tuekam, Shudong Zhang, Berivan Baskin, Gary D Bader, Katerina Michalickova, Tony Pawson, and Christopher WV Hogue. PreBIND and Textomy - mining the biomedical literature for protein-protein interactions using a support vector machine. *BMC Bioinformatics*, 4(11), 2003.

8. Gunes Erkan, Arzucan Ozgur, and Dragomir R. Radev. Semi-supervised classification for extracting protein interaction sentences using dependency parsing. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 228–237, 2007.
9. Mark Girolami and Simon Rogers. Variational bayesian multinomial probit regression with gaussian process priors. *Neural Computation*, 18(8):1790–1817, 2006.
10. Thorsten Joachims. *Advances in Kernel Methods - Support Vector Learning*, chapter Making large-Scale SVM Learning Practical. MIT-Press, 1999.
11. Michael N. Jones, Walter Kintsch, and Douglas J. Mewhort. High-dimensional semantic space accounts of priming. *Journal of Memory and Language*, 55(4):534–552, November 2006.
12. Michael N. Jones and Douglas J. K. Mewhort. Representing word meaning and order information in a composite holographic lexicon. *Psychological Review*, 114:1–37, 2007.
13. J D Kim, T Ohta, Y Tateisi, and J Tsujii. GENIA corpus–semantically annotated corpus for bio-textmining. *Bioinformatics*, 19 Suppl 1:180–182, 2003.
14. M Krallinger, F Leitner, C Rodriguez-Penagos, and A Valencia. Overview of the protein-protein interaction annotation extraction task of biocreative ii. *Genome Biol*, 9 Suppl 2, 2008.
15. T K Landauer, P. W. Foltz, and D Laham. An introduction to latent semantic analysis. *Discourse Processes*, 25:259–284, 1998.
16. David D. Lewis. Naive (Bayes) at forty: The independence assumption in information retrieval. In *ECML '98: Proceedings of the 10th European Conference on Machine Learning*, pages 4–15, London, UK, 1998. Springer-Verlag.
17. K. Lund and C. Burgess. Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior Research Methods, Instrumentation, and Computers*, 28:2003–2208, 1996.
18. Edward M. Marcotte, Ioannis Xenarios, and David Eisenberg. Mining literature for protein-protein interactions. *Bioinformatics*, 17:359 – 363, 2001.
19. Zsolt Minier, Zalan Bodo, and Lehel Csato. Wikipedia-based kernels for text categorization. In *SYNASC '07: Proceedings of the Ninth International Symposium on Symbolic and Numeric Algorithms for Scientific Computing*, pages 157–164, Washington, DC, USA, 2007. IEEE Computer Society.
20. Sebastian Padó and Mirella Lapata. Dependency-based construction of semantic space models. *Comput. Linguist.*, 33(2):161–199, 2007.
21. T. Polajnar and M. Rogers, S. andGirolami. An evaluation of gaussian processes for sentence classification and protein interaction detection. Technical report, University of Glasgow, Department of Computing Science, 2008.
22. S Pyysalo, F Ginter, J Heimonen, J Björne, J Boberg, J Järvinen, and T Salakoski. Bioinfer: a corpus for information extraction in the biomedical domain. *BMC Bioinformatics*, 8:50–50, 2007.
23. S. Rogers and M. Girolami. Multi-class semi-supervised learning with the ϵ- truncated multinomial probit gaussian process. *Journal of Machine Learning Research Workshop and Conference Proceedings*, 1:17–32, 2007.
24. Dawei Song and Peter D.. Bruza. Discovering information flow using a high dimensional conceptual space. In *In Proceedings of ACM SIGIR 2001*, pages 327–333, 2001.
25. C. J. Van Rijsbergen. *Information Retrieval, 2nd edition.* Dept. of Computer Science, University of Glasgow, 1979.