**Application of Lexical Topic Models to Protein Interaction Sentence Prediction**

Tamar Palijanar and Mark Girolami
University of Glasgow :: Department of Computing Science :: Inference Group :: www.dcs.gla.ac.uk/inference

**Introduction**

Purpose: To identify sentences that describe protein-protein interactions (PPis)

Protein-protein interaction extraction is a key application of text mining to biomedical texts. This area of research is strongly motivated by the needs of biologists investigating subcellular processes of unknown function. PPis reported in biomedical journals are detected by large-scale biomedical experiments. The substantial number of interactions results in a large number of publications, combined with the ease of access to the digitised publications provided by the various publisher portals, has increased the number of results made available each day.

Our goal is to automatically identify sentences that describe PPis in biomedical texts as a way of aiding curators and researchers.

**Method**

In order to address the lack of data we propose a novel way of combining labeled and unlabeled data (semi-supervised learning), by integrating semantic information from unsupervised word-word co-occurrence models trained on a large, unannotated corpus of Open Access articles (OAA) and GENIA (Kim et al., 2007). First, we use a similarity measure to collected data. As both models require in vector space, we are able to apply matrix kernels such as kernel functions. Thus, by applying the cosine and RBF kernels, which are often used for text classification, we get valid kernel matrices (K) that can be integrated into kernel classification methods.

- We use the Gaussian process (GP) classifier, a probabilistic analogue to the Support Vector Machine (SVM), which is frequently applied to text data. We have previously shown that it achieves a significantly higher AUC and F-score than the SVM and Naive Bayes on the Aimed data (Polajnar et al., 2009). Another advantage of the GPs is of not having extra parameters, such as the SVM margin parameter.

The labelled data (Y) is used for classifier training and testing, while the semantic information (K) is integrated directly into the kernel K=K+YK−1Y∗.

- Finally, different semantic kernels are combined with uniform weighting of 1 in order to exploit all of their contributions at once using a probabilistic Multiple Kernel Classifier (PMKL) (Damoulas and Girolami, 2008). Why is this approach sensible? How does it improve performance? How does it work?

**Semantic Models**

- **Gene Ontology (GO)**: Each word with a significant co-occurrence is treated as a word type. The semantic dimension corresponding to that word type is used to compute the word’s basis.

**Results**

- **LDA on H3**: H3 is the HAL matrix containing the co-occurrence information between each word and the two clusters to the either side of it. We concentrate only on the words occurring in the training data. For the Aimed dataset, with the stemmed features, the largest single-kernel AUC and F-score were found with the H3 matrix. In order to study the relationships between the words in this matrix we employed HAL. There is a continuous improvement in the LDA-loadedness with an increase in the number of topics; however, even with 40 topics there is a sensible grouping of terms. Each of these topics shows an aspect of the types of protomic and genomic experiments described in this dataset.

- **Figure**: A visualisation of the similarities contained in a BEAGLE matrix created from the BC data and the OAA dataset with DaCoTa.

**Background**

Topic models such as Latent Dirichlet Allocation (LDA) (Blei et al., 2003) and Latent Semantic Analysis (LSA) (Landauer et al., 1998) have been used on a variety of text-based linguistic tasks (Blei et al., 2003; Papadimitriou et al., 2000; Zheng et al., 2006), as well as for other applications (Yuan et al., 2005). Here we examine two related models, Hyper-space Analogue to Language (HAL) (Lund and Burgess, 1998) and Bound Encoding of the Aggregate Language Environment (BEAGLE) (Jones and Mewhort, 2007). Whereas, in LDA and LSA words are generally grouped based on their co-occurrence in similar documents, in HAL and BEAGLE words are grouped based on their co-occurrence with other words. Like LDA, HAL and BEAGLE have been evaluated on a variety of psycho-linguistic tasks such as TOFL word synonym examinations and semantic priming (Jones et al., 2006; Lund and Burgess, 1998). Topic models such as these provide semantic information that is lost through the bag-of-words representation of documents. This property allows us to enrich classification kernels for data-mining applications as PPI sentence classification. Our approach is similar to the semantic smoothing of kernels using WordNet or Wikipedia information (Miner et al., 2007). However, manually constructed ontological lexical information, such as this, is not available for biomedical words. We also gain inspiration from LDA and LSA semantic kernels which often used to smooth kernels based on word-document co-occurrence in the training data (Aservareh, 2008; Craswell et al., 2002).

**Effect on Sentences**

The quality of the output is so high that even with only these four there is an improvement in classification AUC over the original kernel (0.9307 ± 0.0020). Although, this improvement is not as high as using the full HAL matrix.

**References**