Generative vs Discriminative Approaches to Entity Recognition from Label-Deficient Data

Cyril Goutte, Éric Gaussier, Nicola Cancedda, Hervé Déjean

Xerox Research Centre Europe – 6, ch. de Maupertuis – 38240 Meylan – France
{Cyril.Goutte,Eric.Gaussier,Nicola.Cancedda,Herve.Dejean}@xrce.xerox.com

Abstract

Annotating biomedical text for Named Entity Recognition (NER) is usually a tedious and expensive process, while unannotated data is freely available in large quantities. It therefore seems relevant to address biomedical NER using Machine Learning techniques that learn from a combination of labelled and unlabelled data. We consider two approaches: one is discriminative, using Support Vector Machines, the other generative, using mixture models. We compare the two on a biomedical NER task with various levels of annotation, and different similarity measures. We also investigate the use of Fisher kernels as a way to leverage the strength of both approaches. Overall the discriminative approach using standard similarity measures seems to out-perform both the generative approach and the Fisher kernels.

Résumé

L'annotation de textes médicaux pour la reconnaissance d'entités nommées est un travail pénible et couteux, alors que, au contraire, de larges quantités de données non annotées sont disponibles gratuitement sur le Web, par exemple sur MedLine/PubMed. Il est donc particulièrement pertinent de s'attaquer au problème de la reconnaissance d'entités biomédicales à l'aide de techniques d'apprentissage automatique qui apprennent à partir de mélanges de données annotées et non annotées. Nous considérons deux approches: l'une est discriminative, elle repose sur l'utilisation de modèles à point support ou SVM ; l'autre est générative et utilise des modèles de mélange. Nous comparons les deux approches sur une tache de reconnaissance d'entités biomédicales avec divers niveaux d'annotation, et en utilisant différentes mesures de similarité. Nous étudions aussi l'utilisation des noyaux de Fisher afin de tenter de combiner les atouts de chacune des deux approches. Dans l'ensemble, l'approche discriminative utilisant des mesures de similarité standard semble plus performante que l'approche génerative, ainsi que l'utilisation des noyaux de Fisher.

Keywords: biological entity recognition, partially labelled data, discriminative, generative, support vector machines, transductive inference, mixture models, fisher kernels.

1. Introduction

Entity recognition is a crucial step in information extraction. In technical domains such as biomedicine, it is often necessary to recognise specific entities such as protein or gene names. Machine Learning techniques are therefore attractive, as they allow to automatically learn an entity recognition engine for a new domain, with minimal involvement from the user.

One often cited drawback of the Machine Learning approach is that it relies on a corpus, which is usually annotated manually, often a tedious and costly task, especially in technical domains. On the other hand, unannotated data is usually relatively plentiful and freely available, eg in the biomedical domain (cf. http://www.pubmed.org).

As a consequence, it seems particularly relevant to investigate the use of recent Machine Learning approaches that learn from a combination of labelled and unlabelled data, to tackle the

Named Entity Recognition (NER) problem in technical domains. Such approaches include transductive inference (Vapnik, 1995) or learning probabilistic models from partially labelled data (Miller and Uyar, 1997). The former is used in a discriminative setting, estimating directly the labelling of the unlabelled data, while the latter is generative, estimating a probabilistic model that can generate both labelled and unlabelled data. As a dataset containing both labelled and unlabelled data may be seen as a dataset with "missing" labels, it is also called *label-deficient*.

In this contribution, we wish to explore the differences, and the possible links, between these discriminative and generative approaches, in the context of NER. Section 2 introduces the NER problem in the framework of (supervised) categorisation. Section 3 briefly reviews the discriminative and generative approaches, and describes several techniques: transductive inference for Support Vector Machines (3.1.), probabilistic models of label-deficient data (3.2) and Fisher kernels (3.3). In section 4, we explore the experimental results obtained with these approaches on a biological entity recognition task, and we conclude with a discussion.

2. Categorisation for Named Entity Recognition

In our work, we focus on the problem of recognising specific biomedical entities from abstracts of scientific articles. For example, in:

> Inhibition of the activity of Drosophila suppressor of *hairless* and of its human homolog, *KBF2/RBP-J kappa*, by protein interaction.

we wish to recognise that *hairless* and *KBF2/RBP-J kappa* are gene names, while all other terms are irrelevant. We focus on identifying names of genes, proteins or RNA, all other terms (species, chemical names) being irrelevant for this task. Note, however, that we adopt a general Machine Learning approach, which should be applicable to other classes of entities, provided the necessary (small) amount of annotation is available. In addition, we are more interested in the comparison between generative and discriminative approaches using label-deficient data than to the raw performance of either approaches. Several other factors, such as the choice of the appropriate feature set, may influence the final performance of the system.

Many successful approaches to NER formulate the problem in terms of categorisation: is a candidate term an entity (category 1, relevant) or not (category 0, irrelevant)? Examples of this include most of the contributions to the shared task of the last two CoNLL conferences (Roth and van den Bosch, 2002; Daelemans and Osborne, 2003) and several approaches to Biomedical NER (Kazama *et al.*, 2002; Lee *et al.*, 2003; Takeuchi and Collier, 2003; Yamamoto *et al.*, 2003).

In our case, gene, protein and RNA names are relevant, all other terms are not. Given an example $x \in \mathcal{X}$, the entity recognition problem is formulated as the problem of learning a categoriser $h : \mathcal{X} \to \{0; 1\}$, such that the probability that a term is recognised correctly, $P(h(x) = y)$, is maximised over the distribution of (x, y).

3. Learning from partially labelled data

In many text processing applications, it is easy to obtain large amounts of unannotated data, and it may be costly to annotate them. Machine Learning techniques that learn from partially labelled data have therefore been applied to eg text categorisation with some success (Joachims, 1999; Nigam *et al.*, 2000).

The usual way to learn the function h that recognises the relevant entitites is through a sample

of annotated data $(x^{(i)}, y^{(i)})$. It had been argued that having large amounts of non annotated data in the form of additional data $(x^{(j)})$ could in principle help better characterise the classes and therefore improve the categorisation abilities. Imagine for example that the data is formed of a small number of well separated components, each corresponding to a category. The components can be modelled arbitrarily well using unlabelled data alone (possibly in large quantities). Once the components are modelled, little labelled data would suffice to assign a proper label to each of them. If the components are well approximated, the performance would be good, much better than when the classes are learned on the basis of labelled data alone.

This justification is generative in nature, as it relies on the ability to model the data and its labelling. In standard classification problems, however, it has been noted that discriminative methods often out-perform generative techniques, especially in the limit of large sample sizes (Rubinstein and Hastie, 1997; Ng and Jordan, 2002). With partially labelled data, there is also a distinction between discriminative and generative methods. Support Vector Machines are discriminative, and may handle unlabelled data using transductive inference (sec. 3.1.), while probabilistic models of label-deficient data are generative (sec. 3.2.). Fisher kernels (sec. 3.3.) represent a way to bridge the two, by using the generative model to derive a similarity that is used in the discriminative approach.

3.1. *Transductive inference for discriminant analysis*

The standard statistical learning paradigm is to induce a model (eg a Support Vector Machine) from the data and deduce the labelling of test data from this model. Vapnik (1995) argues that the model may actually be superfluous, and advocates the use of *transductive learning* to directly estimate the labelling without first learning the model from the training data. Given an input test data x^*, the transductive approach (Vapnik, 1995; Saunders *et al.*, 1999) tries to estimate directly the decision y^* rather than induce a model $\widehat{h}$ and infer that $y^* = \widehat{h}(x^*)$. Whith little unlabelled data, this is relatively straightforward, but quickly becomes impractical as the number of unlabelled examples increases. Fortunately, using a few heuristics, it is possible to provide an efficient approximate solution to transductive inference with SVM (Joachims, 1999b). In our work, we use this solution, as implemented in Thorsten Joachims' SVMlight software (Joachims, 1999a).

The algorithm starts by labelling the unlabelled data using a SVM trained inductively on the labelled data alone, then repeatedly infer a SVM on the data with the completed labels and tries to improve on the solution by swapping the labels of pairs of examples. This algorithm does converge towards a stable, although obviously not necessarily optimal, solution. Using transductive inference, unlabelled data are taken into account as "test data", and although the inferred labels are usually not used, they have an influence both on the labelling of the true test data, and on the resulting model.

3.2. *Generative models for label-deficient data*

Miller and Uyar (1997) proposed a mixture model to handle combinations of labelled and unlabelled data. Both data and labels are assumed to be generated independently by components of the mixture: $P(x, y) = \sum_\alpha P(\alpha)P(x|\alpha)P(y|\alpha)$. By marginalising over labels y, $P(x) = \sum_\alpha P(\alpha)P(x|\alpha)$. In order to model continuous data, Miller and Uyar (1997) used Gaussian mixtures for the data-dependent component distributions $P(x|\alpha)$. Applications to text processing typically rely on mixtures of binomial distributions. Here, we use a co-occurrence model (Hofmann, 1999) in which each observation x is the co-occurrence of an entity e and a feature f, and $P(x|\alpha) = P(e|\alpha)P(f|\alpha)$. Given a dataset $\mathcal{D}$ composed of labelled and unla-

belled data, $\mathcal{D} = \mathcal{L} \cup \mathcal{U}$, with $\mathcal{L} = \left\{(x^{(i)}, y^{(i)})\right\}_{i=1\ldots N}$ and $\mathcal{U} = \left\{x^{(i)}\right\}_{i=N+1\ldots M}$, parameters are estimated by maximising the (log-)likelihood:

$$L(\mathcal{D}) = \sum_{i=1}^{N} \ln P(x^{(i)}, y^{(i)}) + \sum_{i=N+1}^{M} \ln P(x^{(i)})$$

This may be performed using 2 variants of the Expectation-Maximisation algorithm (Dempster *et al.*, 1977), depending on which latent variables are considered: components alone or components and labels. Both variants maximise the same likelihood and therefore yield similar results. Here we use the former, aka EM1 (Miller and Uyar, 1997), see appendix for details.

Enforcing no constraint on the label generation (so-called *soft partitioning*), especially when classes are unbalanced, results in *cluster impurity*: all or most components contain significant portions of examples from all classes. Components are therefore badly aligned with the discrimination task, and the model yields bad categorisation performance. In order to avoid this issue in the naturally unbalanced problem of NER, we use *hard partitioning*: components are tied to a specific class, ie $P(y|\alpha)$ take binary 0/1 values.

Once the model parameters have been estimated, a label may be assigned to a new example x according to $P(y|x) = (\sum_\alpha P(\alpha)P(x|\alpha)P(y|\alpha))/(\sum_\alpha P(\alpha)P(x|\alpha))$.

3.3. *Fisher kernels*

Jaakkola and Haussler (1999) introduced Fisher kernels as a similarity measure derived from a probabilistic model. This may be useful whenever a probabilistic model of the data exists, as the model-induced similarity may be different from the similarity between observed features. Denoting the log-likelihood for example x by $\ell(x) = \ln P(x|\theta)$, and using the Fisher information matrix $\mathbf{I_F} = \mathsf{E}\left(\nabla\ell(x)\nabla\ell(x)^\top\right)$, the Fisher kernel is:

$$FK(x_1, x_2) = \nabla\ell(x_1)^\top \mathbf{I_F}^{-1} \nabla\ell(x_2) \tag{1}$$

For a suitable choice of parameterisation, $\mathbf{I_F}$ is usually approximated by the identity. In the context of our work there are at least two ways to use Fisher kernels: we can either learn an unsupervised model of the input data $P(x)$ alone, or learn a label-deficient mixture model using the partially annotated data as explained above. In both cases, we then derive a Fisher kernel using equation 1. In our case, these two alternatives correspond to PLSA (Hofmann, 2000) and to the EM1 mixture model. It turns out that the parameters in both models are identical, although they are estimated by maximising different likelihoods, and therefore lead to different parameter estimates. The expression for the Fisher kernels, however, is identical:

$$FK(x_1, x_2) = \sum_\alpha \frac{P(\alpha|e_1)P(\alpha|e_2)}{P(\alpha)} + \sum_f \widehat{P}(f|e_1)\widehat{P}(f|e_2) \sum_\alpha \frac{P(\alpha|f, e_1)P(\alpha|f, e_2)}{P(f|\alpha)} \tag{2}$$

$\widehat{P}(f|e_1)$ (resp. $\widehat{P}(f|e_2)$) is the normalised observed frequency of feature f for entity e_1 (resp. e_2) corresponding to example x_1 (resp. x_2), and all other parameters are obtained during training. Although the expression is identical for PLSA and for EM1, the kernels differ through different parameter estimates. In particular, hard partitioning in EM1 ensures that $FK(x_1, x_2) = 0$ for all pairs of examples with differing labels.

Feature	Value
FULL	1
%LexGENE	1
SYNONYM	1
DICOAMB	1

Feature	Value
LEMME_hairless	1
ADJ	1
LC_Drosophila	1
LC_suppressor	1
LC_PREP	2

Feature	Value
RC_CONJ	1
RC_PREP	1
RC_PRON	1
RC_human	1

Table 1. Features of hairless *in* "of Drosophila suppressor of hairless and of its human"

4. Experimental results

Our dataset was formed using 184 abstracts queried from MedLine. These abstracts were manually annotated by a trained biologist with various biomedical entities. Of these, we focus on gene, protein and RNA names only. We use 122 abstracts as development set (training and validation) and 62 abstracts for testing. All abstracts are tokenised, lemmatised and tagged with part-of-speech information using Xerox linguistic tools. In addition we apply a pre-filtering step that discards all terms that are in a dictionary of common English and are not in the dictionary of possible biological terms. The assumption behind this pre-filtering step is that biological terms that are ambiguous with general English words are well identified and therefore will be in our biomedical resources. Indeed, this filter discards 80% of the original tokens (31617 out of 40398) with a recall of 94% on relevant tokens (2394 out of 2550). We end up with 8781 candidates: 5865 in the development set and 2916 in the test set.

For all candidates, we generate four types of features: spelling (uppercase, digits, etc.), lexical (presence in various dictionaries), linguistic (part-of-speech and lemma) and contextual (words in a 4-word context on either side of the candidate). Table 1 shows all the non-zero features for one of the words in our earlier example. In total, there are 7277 features, although few of them are non-zero for each candidate.

In order to analyse the behaviour of various methods in the presence of labelled and unlabelled data, we retain a variable proportion of the annotation, between 1% and 67%. For each level of partial annotation, we sample 10 different sets of labels, and average the performance over these 10 samples.

In all cases, the baseline is a simple dictionary lookup, combined with part-of-speech information: a candidate is tagged as a relevant entity iff it is present in one of the dictionaries of relevant entities *and* it is tagged as a noun. The precision/recall of this baseline are 58.44%/86.10%, giving a F_1 score of 69.62%.

We first consider Support Vector Machines trained with inductive and transductive inference, and the mixture model with hard partitioning, trained with EM1. Figure 1 shows that transductive inference consistently outperforms both inductive inference and the mixture model. Interestingly, the performance of transductive inference is always better than the baseline, except for one point (RBF kernel at 1% annotation). To put these results into prospective, consider that 2% annotation corresponds to 117 annotated examples (31 positives, 86 negatives) or around $2\frac{1}{2}$ abstracts, ie very little actual annotation work.

All methods yield very similar performance above 16% annotation, but below this level, the performance of transductive inference degrades much more gracefully. Inductive inference and the mixture model need 4 to 8% of annotation to outperform the baseline. Note that although the baseline uses no training data *per se*, it actually relies on a fair amount of prior knowledge on the task and on the available ressources. On the other hand, the Machine Learning method used

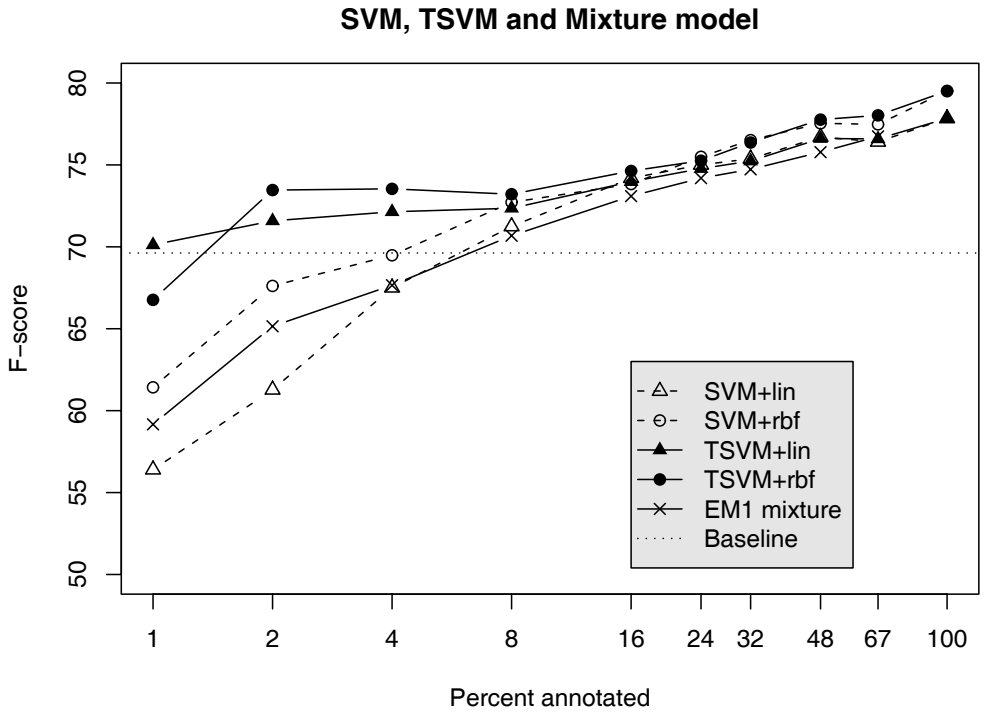

Figure 1. Performance of SVM with lin(ear) and rbf kernels trained using inductive and transductive inference, and mixture model for label-deficient data (EM1). Median F-score over 10 random samples.

here rely only on the available data, and do not use any prior information beside the choice of the feature set. It is therefore not counter-intuitive that the baseline should manage to outperform some of the ML methods when little annotated data is available.

We then address a second relevant question: can we bridge the gap between the discriminative SVM and the generative mixture model using eg Fisher kernels as a similarity measure. In order to investigate this, we calculated the SVM associated with the purely unsupervised PLSA model (FK0) and with the semi-supervised mixture trained by EM1 (FK1). Both kernels are used to train SVM with both inductive and transductive inference. The results are displayed in figure 2. For comparison, we plot the results obtained with the RBF kernel, the best performing of the standard kernels in figure 1.

Clearly the performance of the Fisher kernel derived from PLSA is dire, although transductive inference seems to help. We attribute the poor performance to the problem of cluster impurity, which prevents the model from obtaining components which really correspond to identifiable labels. The performance of the Fisher kernel trained on the label-deficient mixture (FK1) is close to the performance of the RBF kernel, although it is consistently inferior. Finally, FK1 is the only kernel for which transductive inference does not increase, but rather decreases performance.

5. Discussion and conclusion

In this paper, we investigated the use of discriminative versus generative techniques for learning NER from partially labelled data. Based on our experimental results, the discriminative SVM seem to perform significantly and consistently better than the generative mixture model. In fact the mixture model, even though it uses additional unlabelled data, does not manage to

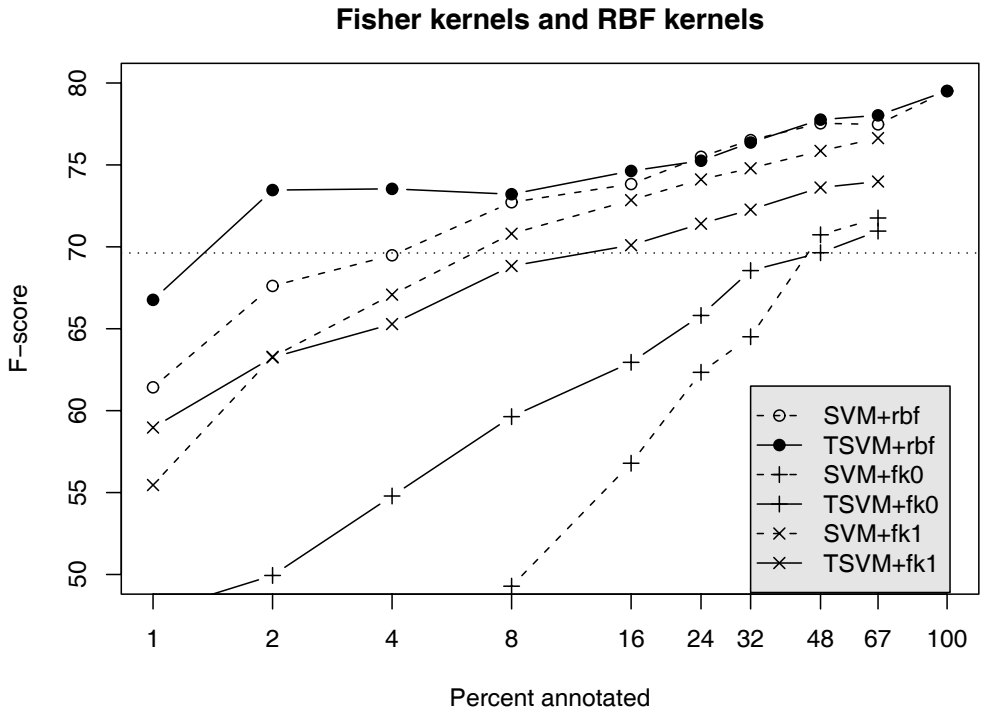

Figure 2. Comparison between Fisher kernels derived from PLSA (FK0), from the semi-supervised mixture (FK1), and standard RBF kernel. Median over 10 random samples.

outperform the SVM trained by inductive inference on the labelled data only. The discriminative approach using transductive inference manages to outperform the baseline using as little as 1% of annotation, ie slightly more than 1 abstract, while all other approaches need at least 4 to 8 times as much data to reach that level of performance. Transductive inference is consistently at least as good, and often significantly better, than inductive inference, for almost all kernels, with the notable exception of the Fisher kernel derived from the generative model.

Overall, the standard RBF kernel performs best. This may be attributed to the fact that it performs an implicit feature expansion into an infinite dimensional space, while all other kernels that we tried (linear and Fisher kernel) use a low-dimensional feature space. It may therefore be easier to "pick" efficient dimensions from the infinite number of implicit RBF dimension, than to work with the relatively few available dimensions in all other cases. The fisher kernels never manage to outperform the standard kernel in our experiments. For FK0, this is mostly due to the *cluster impurity* problem. For FK1, we believe that this is due to the geometry of the implicit feature space when the model uses hard partitioning. In that case, the annotated positive and negative examples belong to two simplexes in two orthogonal subspaces, such that the decision function outside the space spanned by the annotated examples is essentially arbitrary.

These results suggest that using partially labelled data efficiently may yield large performance gains. To illustrate this further, future experiments will apply these techniques to the full annotated dataset, using additional unannotated data queried from PubMed.

Acknowledgements

We thank Anne Schiller, Ágnes Sandor and Violaine Pillet for help with the data. This research was supported by the EC under the KerMIT project (IST-2001-25431).

Appendix: EM equations for label-deficient data

The EM equations for EM1 and hard partitioning are:

$$C^{(t)}(\alpha,i) = \langle P(\alpha|x^{(i)},y^{(i)})\rangle = \frac{P(\alpha)P(x^{(i)}|\alpha)P(y^{(i)}|\alpha)}{\sum_\alpha P(\alpha)P(x^{(i)}|\alpha)P(y^{(i)}|\alpha)} \qquad \text{E-step, labelled data} \qquad (3)$$

$$C^{(t)}(\alpha,i) = \langle P(\alpha|x^{(i)})\rangle = \frac{P(\alpha)P(x^{(i)}|\alpha)}{\sum_\alpha P(\alpha)P(x^{(i)}|\alpha)} \qquad \text{E-step, unlabelled data} \qquad (4)$$

$$P^{(t+1)}(\alpha) = \frac{1}{N}\left(\sum_i C^{(t)}(\alpha,i)\right) \qquad \text{M-step for } P(\alpha) \qquad (5)$$

where $\langle\cdot\rangle$ indicates the expectation conditioned on the data and current parameter estimates. The M-step equations for the co-occurrence model $P(x|\alpha) = P(e|\alpha)P(f|\alpha)$ (Hofmann, 1999) are:

$$P^{(t+1)}(e|\alpha) = \frac{\sum_{i,e^{(i)}=e} C^{(t)}(\alpha,i)}{\sum_i C^{(t)}(\alpha,i)} \text{ and } P^{(t+1)}(f|\alpha) = \frac{\sum_{i,f^{(i)}=f} C^{(t)}(\alpha,i)}{\sum_i C^{(t)}(\alpha,i)} \qquad (6)$$

Parameters are obtained by iterating equations 3 through 6, using a deterministic annealing scheme (Ueda and Nakano, 1995) in order to reduce the sensitivity to initial conditions.

References

Ananiadou S. and Tsujii J. (Eds) (2003). In *Proceedings of the ACL workshop on Natural Language Processing in Biomedicine*.

Daelemans W. and Osborne M. (Eds) (2003).In *Proceedings of the Seventh Conf. on Natural Language Learning*.

Dempster A.P., Laird N.M. and Rubin D.B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B*, vol. (1): 1-38.

Hofmann T. (1999). Probabilistic latent semantic analysis. In *Proceedings of the Fifteenth Conf. on Uncertainty in Artificial Intelligence*: 289-296.

Hofmann T. (2000). Learning the similarity of documents: An information-geometric approach to document retrieval and categorization. In *Advances in Neural Information Processing Systems*, vol. (12).

Jaakkola T. S. and Haussler D. (1999). Exploiting generative models in discriminative classifiers. In *Advances in Neural Information Processing Systems*, vol. (11): 487-493.

Joachims T. (1999a). Making large-scale SVM learning practical. In Schölkopf B., Burges C. and Smola A. (Eds), *Advances in Kernel Methods - Support Vector Learning*. MIT-Press.

Joachims T. (1999b). Transductive inference for text classification using support vector machine. In Bratko I. and Dzeroski S. (Eds), *Machine Learning – Proceedings of the 16th Intl Conf.*: 200-209.

Kazama J., Makino T., Ohta Y. and Tsujii J. (2002). Tuning support vector machines for biomedical named entity recognition. In Johnson S. (Ed.), *Proceedings of the ACL-02 Workshop on Natural Language Processing in the Biomedical Domain*: 1-8.

Lee K.-J., Hwang Y.-S. and Rim H.-C. (2003). Two-phased biomedical NE recognition based on SVMs. In Ananiadou S. and Tsujii J. (Eds), *Proceedings of the ACL workshop on Natural Language Processing in Biomedicine*: 33-40.

Miller D.J. and Uyar H.S. (1997). A mixture of experts classifier with learning based on both labelled and unlabelled data. In *Advances in Neural Information Processing Systems (NIPS*9)*: 571-577.

Ng A.Y. and Jordan M. (2002). On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes. In *Advances in Neural Information Processing Systems*, vol. (14).

Nigam K., McCallum A., Thrun S. and Mitchell T. (2000). Text classification from labeled and unlabeled documents using EM. *Machine Learning*, vol. (39/2-3): 103-134.

Roth D. and van den Bosch A. (Eds) (2002). In *Proceedings of the Sixth Conf. on Natural Language Learning*.

Rubinstein Y. D. and Hastie T. (1997). Discriminative vs informative learning. In *Proceedings of the 3rd Intl Conf. on Knowledge Discovery and Data Mining*: 49-53.

Saunders C., Gammerman A. and Vovk V. (1999). Transduction with confidence and credibility. In Dean T. (Ed.), *Proceedings of the Sixteenth Intl Joint Conf. on Artificial Intelligence*: 722-726.

Takeuchi K. and Collier N. (2003). Bio-medical entity extraction using support vector machines. In Ananiadou S. and Tsujii J. (Eds), *Proceedings of the ACL workshop on Natural Language Processing in Biomedicine*: 57-64.

Ueda N. and Nakano R. (1995). Deterministic annealing variant of the EM algorithm. In *Advances in Neural Information Processing Systems*, vol. (7): 545-552.

Vapnik V.N. (1995). *The Nature of Statistical Learning Theory*. Springer.

Yamamoto K., Kudo T., Konagaya A. and Matsumoto Y. (2003). Protein name tagging for biomedical annotation in text. In Ananiadou S. and Tsujii J. (Eds), *Proceedings of the ACL workshop on Natural Language Processing in Biomedicine*: 65-72.