

Exploiting Feature Hierarchy for Transfer Learning in Named Entity Recognition

Andrew Arnold, Ramesh Nallapati, William W. Cohen
Machine Learning Department
Carnegie Mellon University

ACL HLT
Columbus, Ohio
June 16, 2008

Domain: Biological publications

MOLECULAR AND CELLULAR BIOLOGY, Jan. 1994, p. 373-381
0270-7306/94/\$04.00+0
Copyright © 1994, American Society for Microbiology

Vol. 14, No. 1

The Macrophage Transcription Factor PU.1 Directs Tissue-Specific Expression of the Macrophage Colony-Stimulating Factor Receptor

DONG-ER ZHANG, CHRISTOPHER J. HETHERINGTON, HUI-MIN CHEN, AND DANIEL G. TENEN*

Division of Hematology/Oncology, Department of Medicine, Beth Israel Hospital and Harvard Medical School, Boston, Massachusetts 02115

Received 12 July 1993/Returned for modification 26 August 1993/Accepted 22 September 1993

The macrophage colony-stimulating factor (M-CSF) receptor is expressed in a tissue-specific fashion from two distinct promoters in monocytes/macrophages and the placenta. In order to further understand the transcription factors which play a role in the commitment of multipotential progenitors to the monocyte/macrophage lineage, we have initiated an investigation of the factors which activate the M-CSF receptor very early during the monocyte differentiation process. Here we demonstrate that the human monocytic M-CSF receptor promoter directs reporter gene activity in a tissue-specific fashion. Since one of the few transcription factors which have been implicated in the regulation of monocyte genes is the macrophage- and B-cell-specific PU.1 transcription factor, we investigated whether PU.1 binds and activates the M-CSF receptor promoter. Here we demonstrate that both *in vitro*-translated PU.1 and PU.1 from nuclear extracts bind to a specific site in the M-CSF receptor promoter just upstream from the major transcription initiation site. Mutations in this site which eliminate PU.1 binding decrease M-CSF receptor promoter activity significantly in macrophage cell lines only. Furthermore, PU.1 transactivates the M-CSF receptor promoter in nonmacrophage cells. These results suggest that PU.1 plays a major role in macrophage gene regulation and development by directing the expression of a receptor for a key macrophage growth factor.

The 150-kDa receptor of the macrophage colony-stimulating factor (M-CSF) receptor is encoded by the *c-fms* proto-oncogene, the cellular counterpart of the Susan McDonough strain of the feline sarcoma virus *v-fms* gene (26). The

direct M-CSF receptor tissue-specific expression in the two different tissues (29). In monocytes, transcription initiates at multiple sites immediately upstream of the start codon ATG.

Problem: Protein-name extraction

MOLECULAR AND CELLULAR BIOLOGY, Jan. 1994, p. 373-381
0270-7306/94/\$04.00+0
Copyright © 1994, American Society for Microbiology

Vol. 14, No. 1

The Macrophage Transcription Factor PU.1 Directs Tissue-Specific Expression of the Macrophage Colony-Stimulating Factor Receptor

DONG-ER ZHANG, CHRISTOPHER J. HETHERINGTON, HUI-MIN CHEN, AND DANIEL G. TENEN*

Division of Hematology/Oncology, Department of Medicine, Beth Israel Hospital and Harvard Medical School, Boston, Massachusetts 02115

Received 12 July 1993/Returned for modification 26 August 1993/Accepted 22 September 1993

The **macrophage colony-stimulating factor (M-CSF)** receptor is expressed in a tissue-specific fashion from two distinct promoters in monocytes/macrophages and the placenta. In order to further understand the transcription factors which play a role in the commitment of multipotential progenitors to the monocyte/macrophage lineage, we have initiated an investigation of the factors which activate the **M-CSF receptor** very early during the monocyte differentiation process. Here we demonstrate that the human monocytic M-CSF receptor promoter directs reporter gene activity in a tissue-specific fashion. Since one of the few transcription factors which have been implicated in the regulation of monocyte genes is the macrophage- and B-cell-specific **PU.1 transcription factor**, we investigated whether PU.1 binds and activates the M-CSF receptor promoter. Here we demonstrate that both in vitro-translated PU.1 and PU.1 from nuclear extracts bind to a specific site in the M-CSF receptor promoter just upstream from the major transcription initiation site. Mutations in this site which eliminate PU.1 binding decrease M-CSF receptor promoter activity significantly in macrophage cell lines only. Furthermore, PU.1 transactivates the M-CSF receptor promoter in nonmacrophage cells. These results suggest that PU.1 plays a major role in macrophage gene regulation and development by directing the expression of a receptor for a key macrophage growth factor.

The 150-kDa receptor of the macrophage colony-stimulating factor (M-CSF) receptor is encoded by the *c-fms* proto-oncogene, the cellular counterpart of the Susan McDonough strain of the feline sarcoma virus *v-fms* gene (26). The

direct M-CSF receptor tissue-specific expression in the two different tissues (29). In monocytes, transcription initiates at multiple sites immediately upstream of the start codon ATG.

The Problem

- What we are able to do:
 - Train on large, labeled data sets drawn from same distribution as testing data
- What we would like to be able do:
 - Leverage large, previously labeled data from a **related** domain
 - Transfer learning:
 - Domain we're interested in (data scarce): *Target*
 - Related domain (with lots of data): *Source*
- How we plan to do it:
 - Isolate features with similar distributions across domains
 - Use feature space's inherent **structure** to find these similarities
 - Spread this information using carefully constructed **priors**

Motivation

- Why is transfer important?
 - Often we violate non-transfer assumption without realizing. How much data is truly identically distributed (the *i.d.* from *i.i.d.*)?
 - E.g. Different authors, annotators, time periods, sources
 - Large amounts of labeled data/trained classifiers already exist
 - Why waste data & computation?
 - Can learning be made easier by leveraging related domains/problems?
 - Life-long learning
- Why is structure important?
 - Need some bias as to how different domains' features relate to one another
- Why are priors important?
 - Small bits of selective knowledge
 - Guide learning algorithms
 - Still relatively inexpensive

What we are able to do:

- Supervised learning
 - Train on large, labeled data sets drawn from same distribution as testing data
 - Well studied problem

Training data:

Dear William,
Can you please tell Richard to give Bin the notes from class?
Thanks,
Andrew

Test:

Hi Carol!
Please let me know when you and Jim are meeting for lunch.
Thanks!
--aa

Train:

The neuronal cyclin-dependent kinase p35/cdk5 comprises a catalytic subunit (cdk5) and an activator subunit (p35)

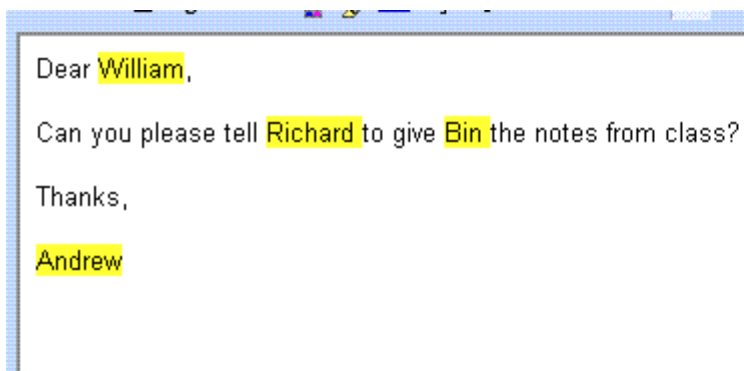
Test:

Reversible histone acetylation changes the chromatin structure and can modulate gene transcription. Mammalian histone deacetylase 1 (HDAC1)

What we would like to be able to do:

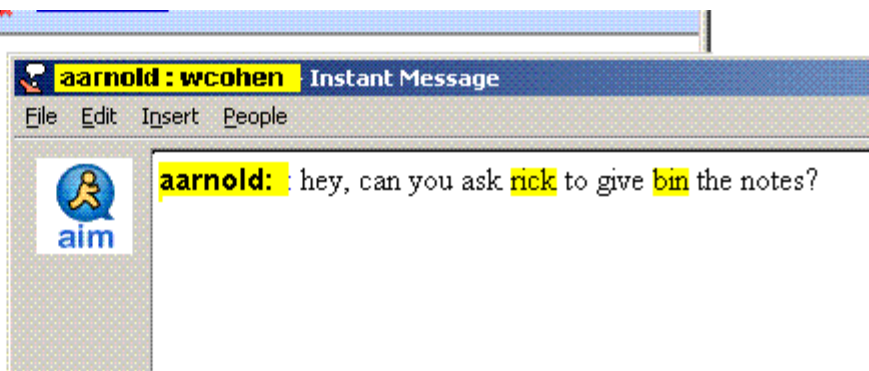
- Transfer learning (domain adaptation):
 - Leverage large, previously labeled data from a **related** domain
 - Related domain we'll be training on (with lots of data): *Source*
 - Domain we're interested in and will be tested on (data scarce): *Target*
 - [Ng '06, Daumé '06, Jiang '06, Blitzer '06, Ben-David '07, Thrun '96]

Train (*source domain*: E-mail):



Dear William,
Can you please tell Richard to give Bin the notes from class?
Thanks,
Andrew

Test (*target domain*: IM):



aarnold:wcohen Instant Message
File Edit Insert People
aarnold: hey, can you ask rick to give bin the notes?

Train (*source domain*: Abstract):

The neuronal cyclin-dependent kinase p35/cdk5 comprises a catalytic subunit (cdk5) and an activator subunit (p35)

Test (*target domain*: Caption):

Neuronal cyclin-dependent kinase p35/cdk5 (Fig 1, a) comprises a catalytic subunit (cdk5, left panel) and an activator subunit (p35, fmi #4)

What we'd like to be able to do:

- Transfer learning (multi-task):
 - Same domain, but slightly different task
 - Related task we'll be training on (with lots of data): *Source*
 - Task we're interested in and will be tested on (data scarce): *Target*
 - [Ando '05, Sutton '05]

Train (*source task*: Names):

Dear William,
Can you please tell Richard to give Bin the notes from class?
Thanks,
Andrew

Test (*target task*: Pronouns):

Hi Carol!
Please let me know when you and Jim are meeting for lunch.
Thanks!
--aa

Train (*source task*: Proteins):

The neuronal cyclin-dependent kinase p35/cdk5 comprises a catalytic subunit (cdk5) and an activator subunit (p35)

Test (*target task*: Action Verbs):

Reversible histone acetylation changes the chromatin structure and can modulate gene transcription. Mammalian histone deacetylase 1 (HDAC1)

The Features

[Class: NEG 1.0] Span 'death'= tokens 120:121 in 536_98374313_9707608_genia_1480.txt/536_98374313_9707608_genia_1480.txt

[Class: NEG 1.0] Span 'domain'= tokens 121:122 in 536_98374313_9707608_genia_1480.txt/536_98374313_9707608_genia_1480.txt

[Class: NEG 1.0] Span 'interacting'= tokens 122:123 in 536_98374313_9707608_genia_1480.txt/536_98374313_9707608_genia_1480.txt

[Class: NEG 1.0] Span 'protein'= tokens 123:124 in 536_98374313_9707608_genia_1480.txt/536_98374313_9707608_genia_1480.txt

[Class: protUnique 1.0] Span 'TRADD'= tokens 124:125 in 536_98374313_9707608_genia_1480.txt/536_98374313_9707608_genia_1480.txt

Features **Source** **Subpopulation**

to TNFRI in associating with the TNFRI death domain interacting protein TRADD .
TNFRI has been recently shown to activate NF

Features **Source** **Subpopulation**

Class label: [ClassLabel: {NEG=1.0}]

Feature Name	Weight
previousLabel.1.NEG	1.0
tokens.eq.charTypePattern.x+	1.0
tokens.eq.lc.protein	1.0
left.tokenNeg_1.eq.charTypePattern.x+	1.0
left.tokenNeg_1.eq.lc.interacting	1.0
left.tokenNeg_2.eq.charTypePattern.x+	1.0
left.tokenNeg_2.eq.lc.domain	1.0
left.tokenNeg_3.eq.charTypePattern.x+	1.0
left.tokenNeg_3.eq.lc.death	1.0
right.token_0.eq.charTypePattern.X+	1.0
right.token_0.eq.lc.tradd	1.0
right.token_1.eq.charTypePattern..	1.0
right.token_1.eq.lc..	1.0
right.token_2.eq.charTypePattern.X+	1.0
right.token_2.eq.lc.tnfri	1.0

Feature Hierarchy

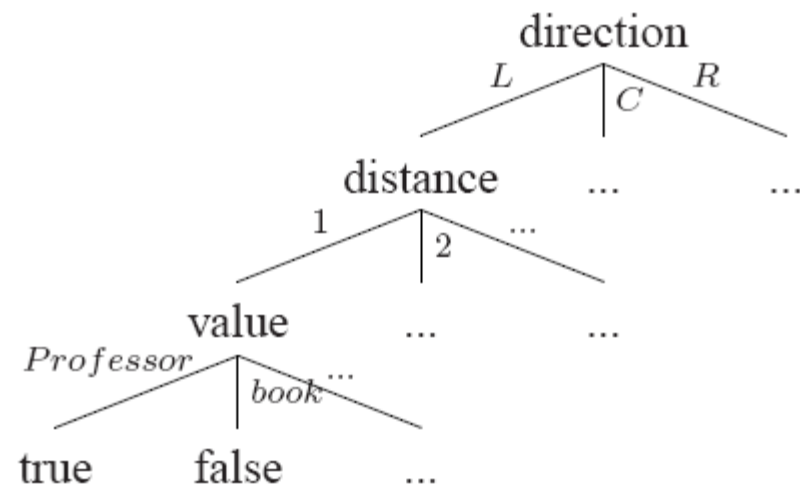
Sample sentence:

Give the book to Professor Caldwell

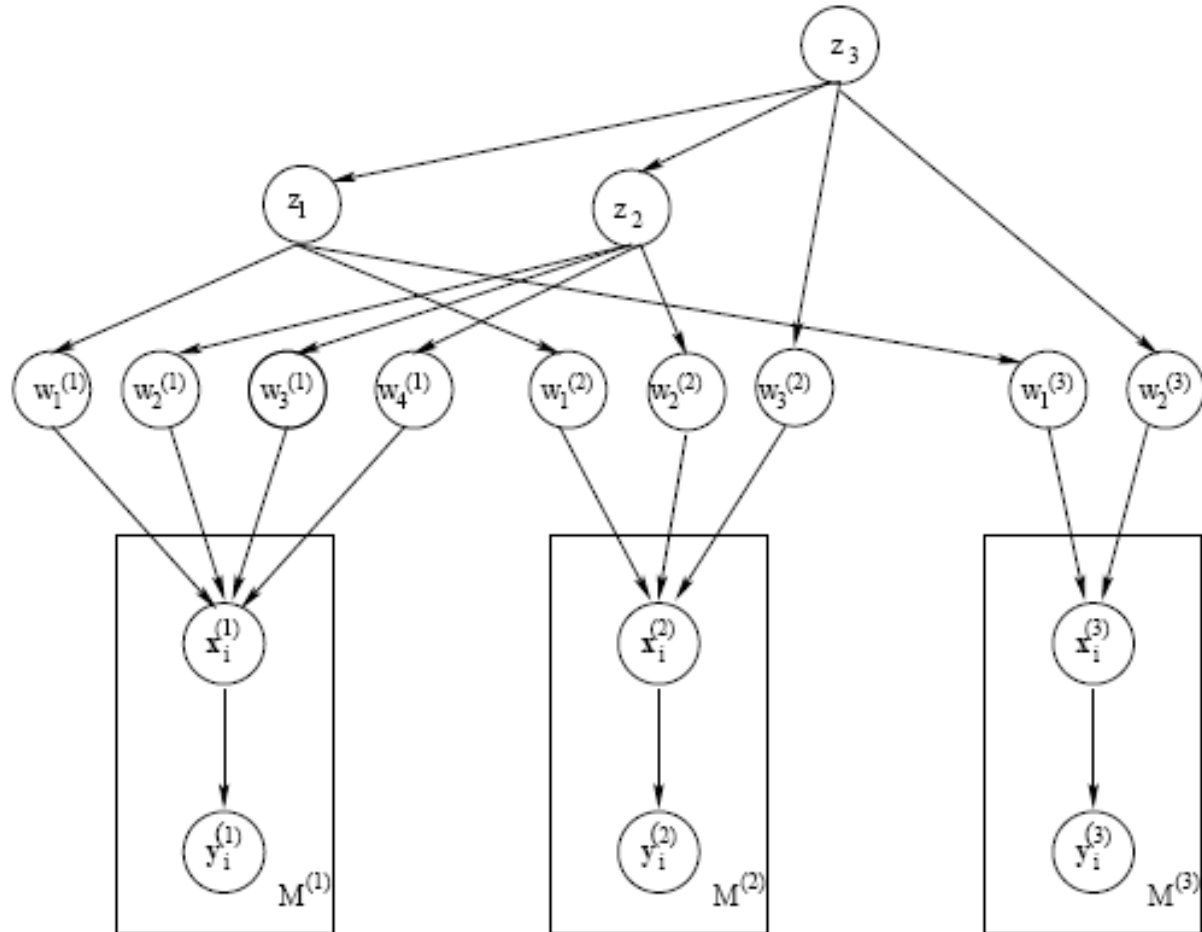
Examples of the feature hierarchy:

LeftToken.*
LeftToken.IsWord.*
LeftToken.IsWord.IsTitle.*
LeftToken.IsWord.IsTitle.equals.*
LeftToken.IsWord.IsTitle.equals.mr

Hierarchical feature tree for 'Caldwell':



Hierarchical prior model (HIER)



- Top level: \mathbf{z} , hyperparameters, linking related features
- Mid level: \mathbf{w} , feature weights per each domain
- Low level: \mathbf{x} , \mathbf{y} , training data:label pairs for each domain

Hierarchical prior model (cont.)

Conditional likelihood of data:

$$P(\mathbf{y}|\mathbf{x}, \mathbf{w}, \mathbf{z}) =$$

Likelihood of data in each domain,
given domain's model parameters:

$$\left\{ \prod_{d=1}^D \prod_{k=1}^{M_d} P(\mathbf{y}_k^{(d)} | \mathbf{x}_k^{(d)}, \Lambda^{(d)}) \right\}$$

Likelihood of each model parameter
in each domain's given its parent's
hyperparameter:

$$\times \left\{ \prod_{d=1}^D \prod_{f=1}^{F_d} \mathcal{N}(\lambda_f^{(d)} | z_{\text{pa}(f^{(d)})}, 1) \right\}$$

Hyperparameters (without leaf nodes):

$$\times \left\{ \prod_{n \in \mathcal{T}_{\text{nonleaf}}} \mathcal{N}(z_n | z_{\text{pa}(n)}, 1) \right\}$$

Approximate algorithm & smoothing

Input: $\mathcal{D}^{source} = (X_{train}^{source}, Y_{train}^{source})$
 $\mathcal{D}^{target} = (X_{train}^{target}, Y_{train}^{target})$;
Feature sets $\mathcal{F}^{source}, \mathcal{F}^{target}$;
Feature Hierarchies $\mathcal{H}^{source}, \mathcal{H}^{target}$
Minimum membership size M

Smoothing feature weights across entire tree can lead to over-smoothing

- Joining unrelated features/domains

Train CRF using \mathcal{D}^{source} to obtain feature weights Λ^{source}

For each feature $f \in \mathcal{F}^{target}$

Initialize: node $n = f$

While ($n \notin \mathcal{H}^{source}$

or $|\text{Leaves}(\mathcal{H}^{source}(n))| \leq M$)

and $n \neq \text{root}(\mathcal{H}^{target})$

$n \leftarrow \text{Pa}(\mathcal{H}^{target}(n))$

Compute μ_f and σ_f using the sample

$\{\lambda_i^{source} \mid i \in \text{Leaves}(\mathcal{H}^{source}(n))\}$

Train Gaussian prior CRF using \mathcal{D}^{target} as data and $\{\mu_f\}$ and $\{\sigma_f\}$ as Gaussian prior parameters.

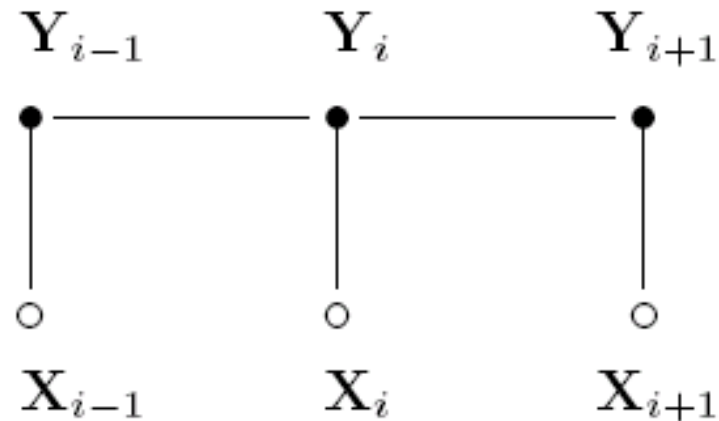
Output: Parameters of the new CRF Λ^{target} .

Instead, can adjust *level* of tree to smooth over

- Also minimum membership size (M)

Models

- Conditional random field (CRF):
 - Sequentially classify tokens, given context
 - Breaks normal i.i.d. assumption
 - Neighbors' predicted class can influence my class



$$p_{\Lambda}(\mathbf{Y} = \mathbf{y} | \mathbf{x}) = \frac{1}{Z(\mathbf{x})} \exp\left(\sum_{i=1}^n \sum_{j=1}^F f_j(\mathbf{x}, y_i) \lambda_j\right)$$

Regularized models

- CRF with Gaussian prior (**GAUSS**):

$$\operatorname{argmax}_{\Lambda} \sum_{k=1}^N \left(\log p_{\Lambda}(\mathbf{y}_k | \mathbf{x}_k) \right) - \beta \sum_j^F \frac{(\lambda_j - \mu_j)^2}{2\sigma_j^2}$$

- Instead of regularizing towards zero
 - Learn model Λ 's on source data
 - During target training
 - Regularize towards source-trained Λ 's (**CHELBA**)

$$\operatorname{argmax}_{\Lambda^{target}} p_{\Lambda^{target}}(Y|X) - \beta ||\Lambda^{target} - \Lambda^{source}||$$

Data

Corpus	Genre	Task
UTexas	Bio	Protein
Yapex	Bio	Protein
MUC6	News	Person
MUC7	News	Person
CSPACE	E-mail	Person

<prot> p38 stress-activated protein kinase
</prot> inhibitor reverses <prot> bradykinin B(1)
receptor </prot>-mediated component of
inflammatory hyperalgesia.

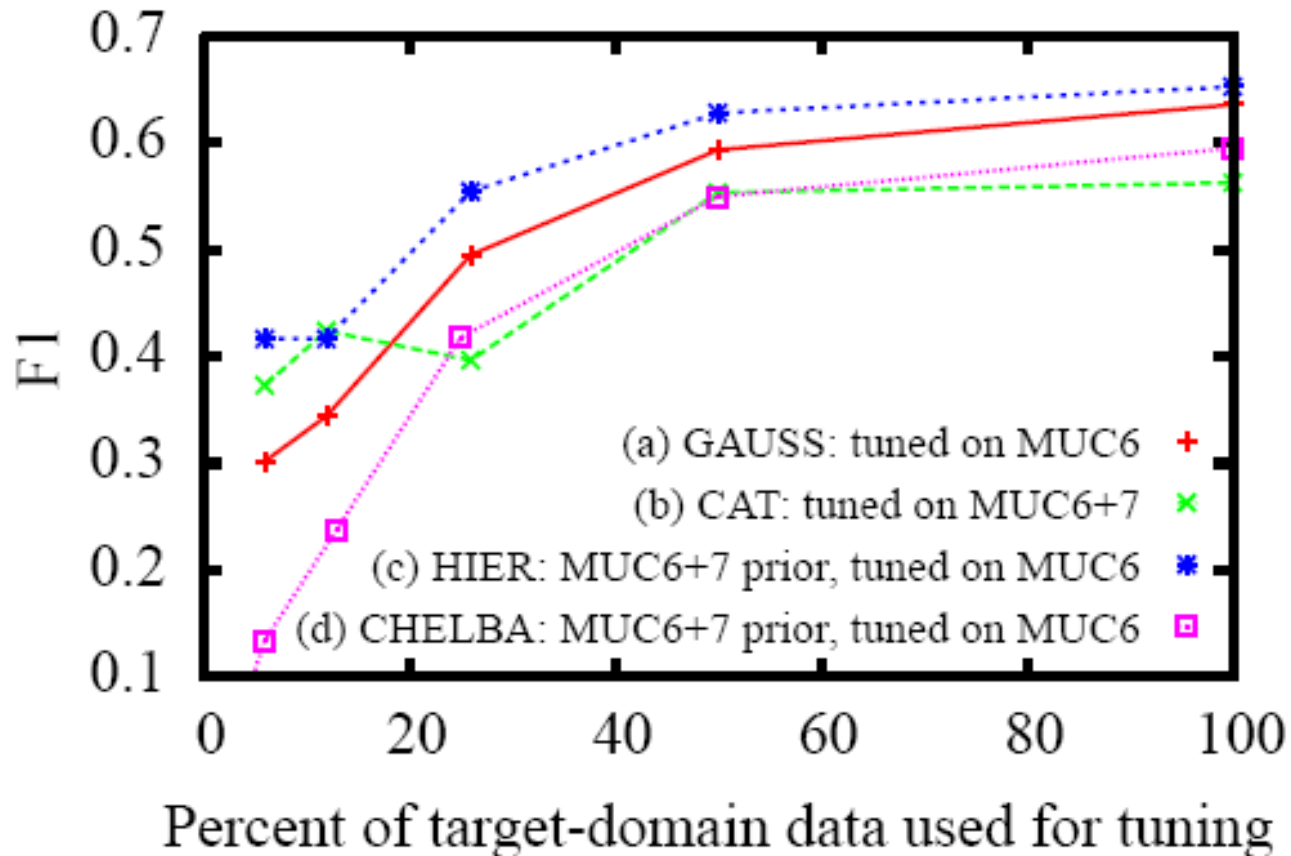
<Protname>p35</Protname>/<Protname>cdk5
</Protname> binds and phosphorylates
<Protname>beta-catenin</Protname> and
regulates <Protname>beta-catenin </Protname> /
<Protname>presenilin-1</Protname> interaction.

- Corpora come from three **genres**:
 - Biological journal abstracts
 - News articles
 - Personal e-mails
- Two **tasks**:
 - Protein names in biological abstracts
 - Person names in news articles and e-mails
- Variety of genres and tasks allows us to:
 - evaluate each method's ability to generalize across and incorporate information from a wide variety of domains, genres and tasks

Experiments

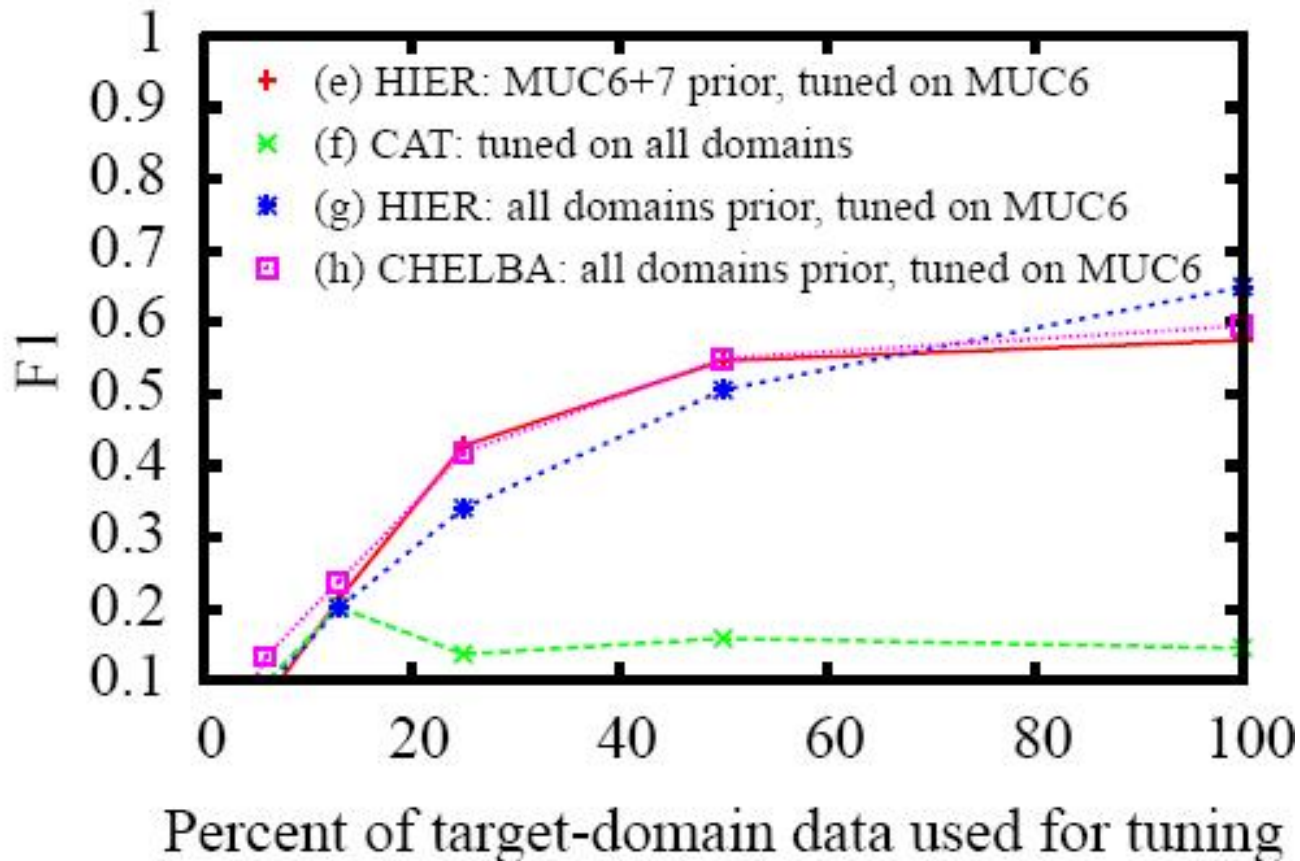
- Compared HIER against three baselines:
 - **GUASS**: CRF tuned on single domain's data
 - Standard $N(0,1)$ prior
 - **CAT**: CRF tuned on concatenation of multiple domains' data, using standard $N(0,1)$ prior
 - **CHELBA**: CRF model tuned on one domain's data, using prior trained on different, related domain's data
- Since few true positives, focused on:
$$F1 := (2 * Precision * Recall) / (Precision + Recall)$$

Results: Intra-genre, same-task transfer



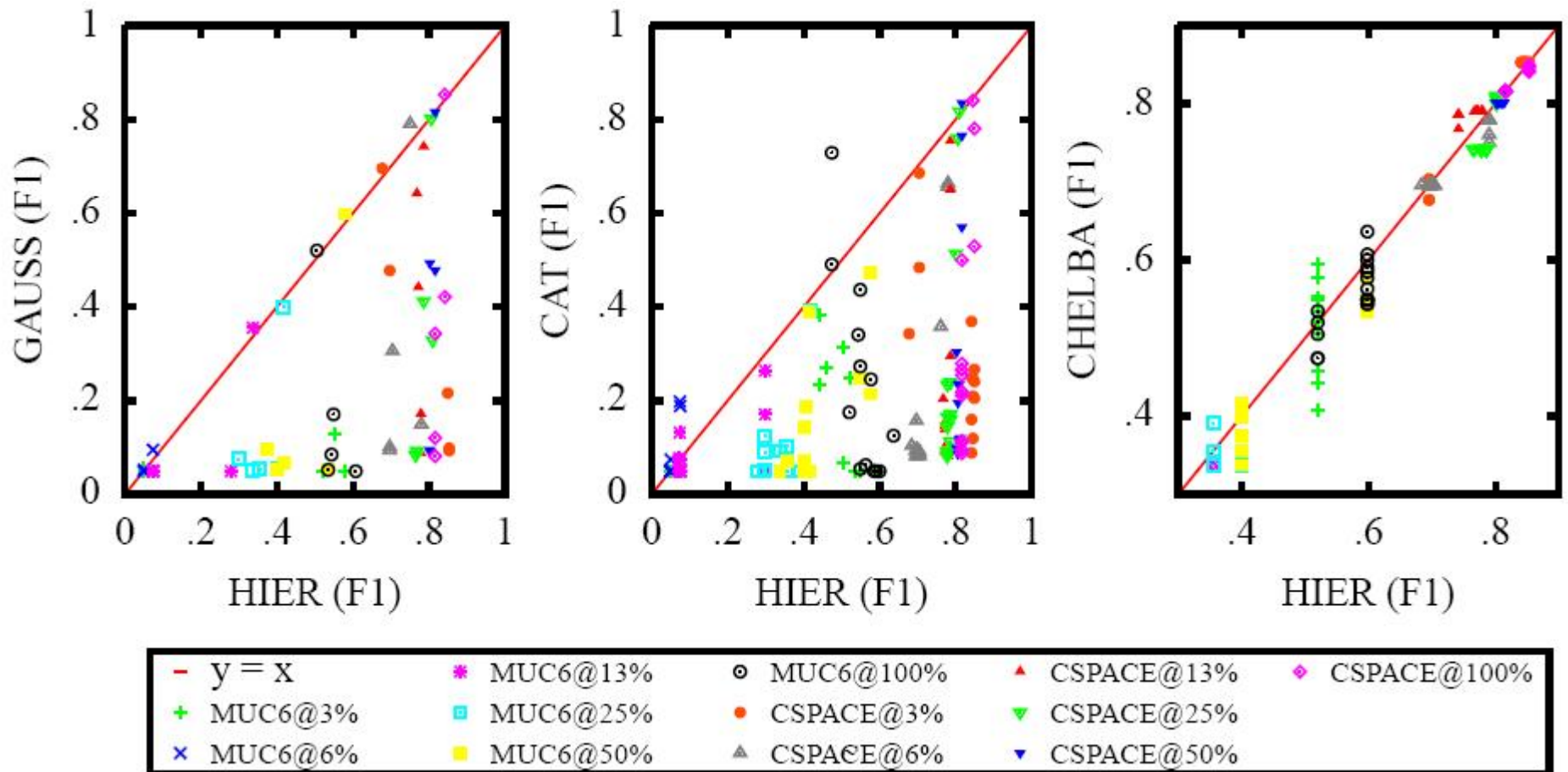
- Adding relevant HIER prior helps compared to GAUSS ($c > a$)
- Simply CAT'ing or using CHELBA can hurt ($d \approx b < a$)
- And never beat HIER ($c > b \approx d$)

Results: Inter-genre, multi-task transfer



- Transfer-aware priors CHELBA and HIER filter irrelevant data
- Adding irrelevant data to priors doesn't hurt ($e \approx g \approx h$)
- But simply CAT'ing it is disastrous ($f \ll e$)

Results: Baselines vs. HIER



- Points **below** $Y=X$ indicate HIER outperforming baselines
 - HIER dominates non-transfer methods (GUASS, CAT)
 - Closer to non-hierarchical transfer (CHELBA), but still outperforms

Conclusions & Future work

- Hierarchical feature priors successfully
 - exploit structure of many different natural language feature spaces
 - while allowing flexibility (via smoothing) to transfer across various distinct, but related domains, genres and tasks
- Future work extends these methods to the semi-supervised and unsupervised settings
- Exploits structure not only in features space, but also in data space
 - E.g.: Transfer from abstracts to captions of papers
From Headers to Bodies of e-mails

☺ iThank you! ☺

¿ Questions ?

References

- Rie K. Ando and Tong Zhang. 2005. A framework for learning predictive structures from multiple tasks and unlabeled data. In *JMLR 6*, pages 1817 – 1853.
- Andrew Arnold, Ramesh Nallapati, and William W. Cohen. 2007. A comparative study of methods for transductive transfer learning. In *Proceedings of the IEEE International Conference on Data Mining (ICDM) 2007 Workshop on Mining and Management of Biological Data*.
- Jonathan Baxter. 1997. A Bayesian/information theoretic model of learning to learn via multiple task sampling. *Machine Learning*, 28(1):7–39.
- Shai Ben-David, John Blitzer, Koby Crammer, and Fernando Pereira. 2007. Analysis of representations for domain adaptation. In *NIPS 20*, Cambridge, MA. MIT Press.
- John Blitzer, Ryan McDonald, and Fernando Pereira. 2006. Domain adaptation with structural correspondence learning. In *EMNLP*, Sydney, Australia.
- A. Borthwick, J. Sterling, E. Agichtein, and R. Grishman. 1998. NYU: Description of the MENE named entity system as used in MUC-7.
- R. Bunescu, R. Ge, R. Kate, E. Marcotte, R. Mooney, A. Ramani, and Y. Wong. 2004. Comparative experiments on learning information extractors for proteins and their interactions. In *Journal of AI in Medicine. Data from ftp://ftp.cs.utexas.edu/pub/mooney/bio-data/proteins.tar.gz*.
- Rich Caruana. 1997. Multitask learning. *Machine Learning*, 28(1):41–75.
- Ciprian Chelba and Alex Acero. 2004. Adaptation of maximum entropy capitalizer: Little data can help a lot. In Dekang Lin and Dekai Wu, editors, *EMNLP 2004*, pages 285–292. ACL.
- S. Chen and R. Rosenfeld. 1999. A gaussian prior for smoothing maximum entropy models.
- William W. Cohen. 2004. Minorthird: Methods for identifying names and ontological relations in text using heuristics for inducing regularities from data. <http://minorthird.sourceforge.net>.
- Hal Daumé III and Daniel Marcu. 2006. Domain adaptation for statistical classifiers. In *Journal of Artificial Intelligence Research 26*, pages 101–126.
- Hal Daumé III. 2007. Frustratingly easy domain adaptation. In *ACL*.
- David Fisher, Stephen Soderland, Joseph McCarthy, Fangfang Feng, and Wendy Lehnert. 1995. Description of the UMass system as used for MUC-6.
- Kristofer Franzén, Gunnar Eriksson, Fredrik Olsson, Lars Asker, Per Lidn, and Joakim Cöster. 2002. Protein names and how to find them. In *International Journal of Medical Informatics*.
- Yves Grandvalet and Yoshua Bengio. 2005. Semi-supervised learning by entropy minimization. In *CAP*, Nice, France.
- Jing Jiang and ChengXiang Zhai. 2006. Exploiting domain structure for named entity recognition. In *Human Language Technology Conference*, pages 74 – 81.
- R. Kraut, S. Fussell, F. Lerch, and J. Espinosa. 2004. Coordination in teams: evidence from a simulated management game.
- John Lafferty, Andrew McCallum, and Fernando Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proc. 18th International Conf. on Machine Learning*, pages 282–289. Morgan Kaufmann, San Francisco, CA.
- S.-I. Lee, V. Chatalbashev, D. Vickrey, and D. Koller. 2007. Learning a meta-level prior for feature relevance from multiple related tasks. In *Proceedings of International Conference on Machine Learning (ICML)*.
- Einat Minkov, Richard C. Wang, and William W. Cohen. 2005. Extracting personal names from email: Applying named entity recognition to informal text. In *HLT/EMNLP*.
- Rajat Raina, Andrew Y. Ng, and Daphne Koller. 2006. Transfer learning by constructing informative priors. In *ICML 22*.
- Bernhard Schölkopf, Florian Steinke, and Volker Blanz. 2005. Object correspondence as a machine learning problem. In *ICML '05: Proceedings of the 22nd international conference on Machine learning*, pages 776–783, New York, NY, USA. ACM.
- Charles Sutton and Andrew McCallum. 2005. Composition of conditional random fields for transfer learning. In *HLT/EMNLP*.
- M. Szafranski, Y. Grandvalet, and P. Morizet-Mahoudeaux. 2007. Hierarchical penalization. In *Advances in Neural Information Processing Systems 20*. MIT press.
- B. Taskar, M.-F. Wong, and D. Koller. 2003. Learning on the test data: Leveraging ‘unseen’ features. In *Proc. Twentieth International Conference on Machine Learning (ICML)*.
- Sebastian Thrun. 1996. Is learning the n -th thing any easier than learning the first? In *NIPS*, volume 8, pages 640–646. MIT.
- J. Zhang, Z. Ghahramani, and Y. Yang. 2005. Learning multiple related tasks using latent independent component analysis.