Sentiment Analysis in new turfs

Balamurali A R,

Aditya Joshi,

Pushpak Bhattacharyya

Sentiment Analysis Group

CFILT, IIT Bombay

Road map

New Domain

New Language

New Text Form

Road map

New Domain

- Introduction
- Problem Statement
- Intuition
- Approach
- System Architecture
- Results

New Language

New Text Form

Sentiment Analysis for a new domain

Based on our work titled ‘The Power of Many and the Ingenuity of the Master: A Novel Approach for Cross-domain Sentiment Tagging’ to be presented at International Conference on Natural Language Processing (ICON) 2010 to be held in December 2010

Introduction (1/2)

- Sentiment Classification
  - Classification of document based on sentiment content
    - Content being positive/negative to a user evaluating the text
  - Supervised and unsupervised approaches exist
  - Supervised approaches perform better
    - Need of labeled data

Supervised Approaches: Pang & Lee(2002), Matsumoto et.al(2005)

Unsupervised Approaches: Turney (2002), Wan(2008)

Introduction (2/2)

- Labeled data for a sentiment classifier
  - – Tedious and expensive process
  - – Domain specific nature of Sentiment Analysis
  - – Utility span of labeled data is limited

Product review domain v/s Movie review domain

Problem we address

- A sentiment classifier trained on one domain performs with a lesser accuracy than on a different domain (Aue and Gamon 2005)

- **Solution**: Cross-Domain Sentiment Adaptation
  - A procedure to use labeled data in an existing domain and use them to create labeled data for new target domains having little or no training data the with help of few hand labeled target data

Intuition

You evolve using the information attained, disregarding the unwanted

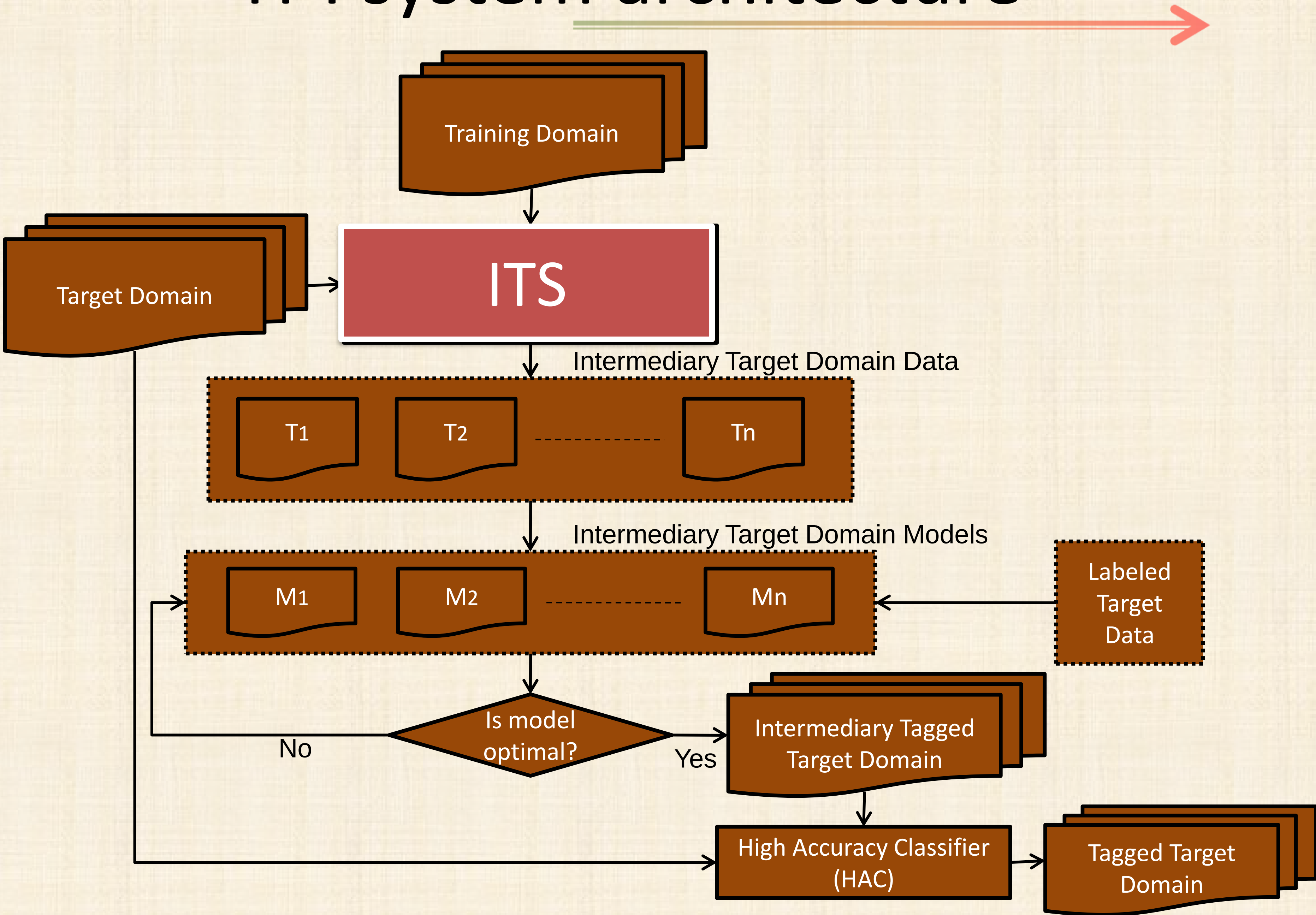
Approach

- Generate noisy tagged data for the target domain using an Meta-classifier trained on a source domain

- Select the *highly probable correct instances*

- Use them to create a high accuracy in-domain classifier
  - This classifier then completely classifies the target domain

H-I system architecture



Training Domain

Target Domain

ITS

Intermediary Target Domain Data

T1

T2

Tn

Intermediary Target Domain Models

M1

M2

Mn

Labeled Target Data

Is model optimal?

No

Yes

Intermediary Tagged Target Domain

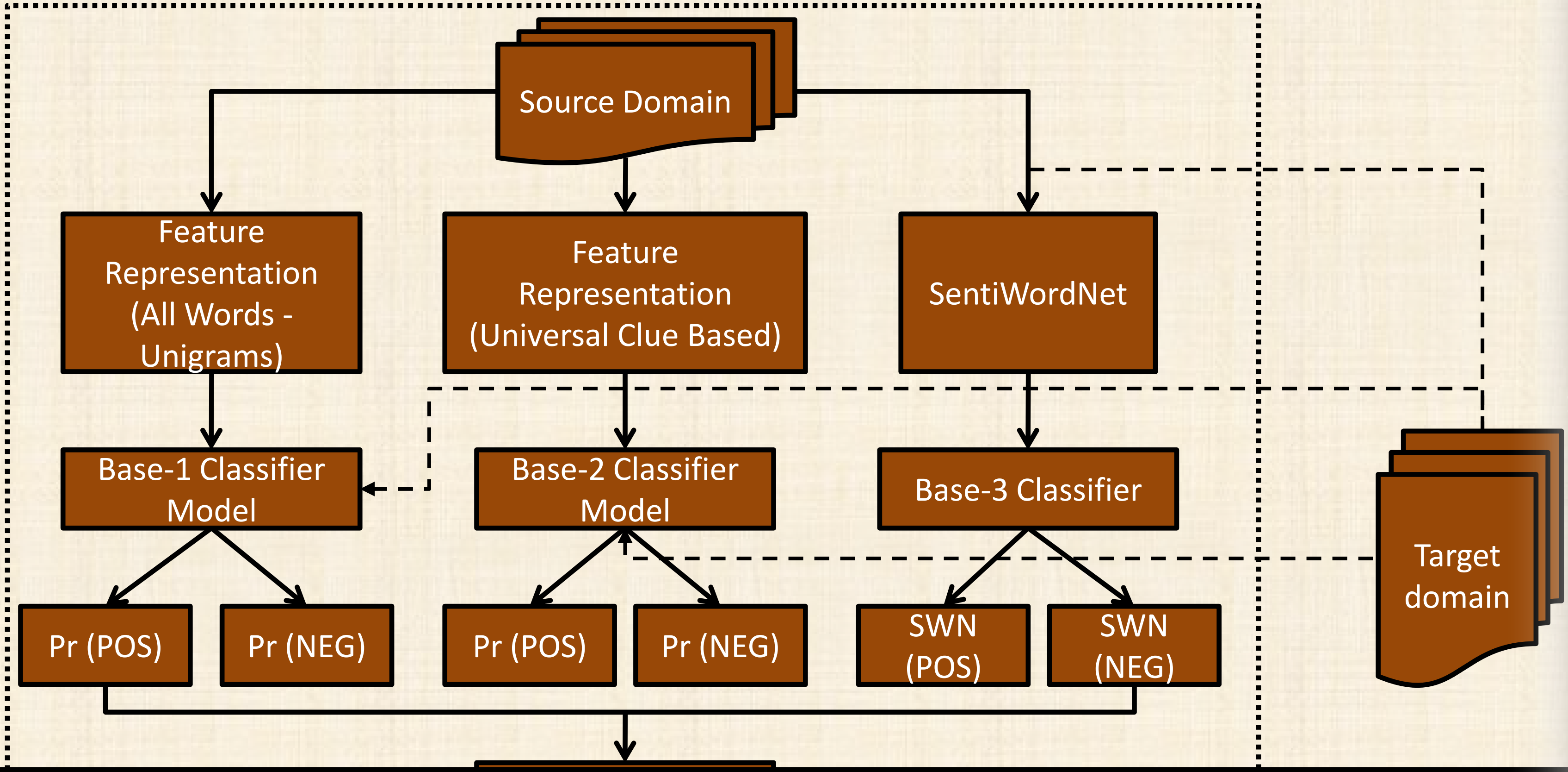
High Accuracy Classifier (HAC)

Tagged Target Domain

Lexicons used

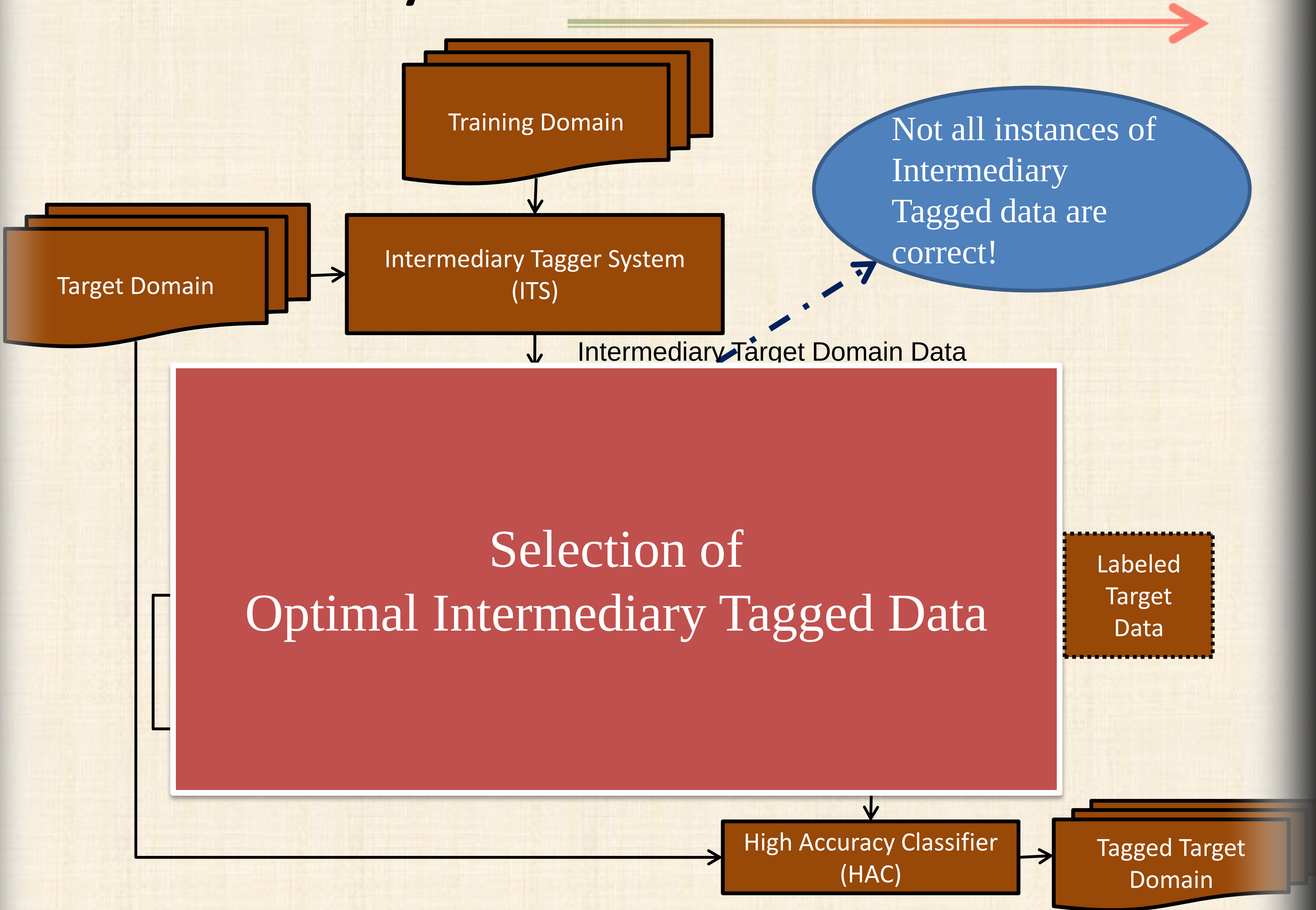
- SentiWordnet (Esuli and Sebastiani, 2006)
  - Attaches sentiment score to each synset
    - Positive , negative & objective score
    - Scores sum up to 1
    - E.g. – *{small}* – *“slight or limited”*
      - *pos – 0.125, neg – 0.125 , obj – 0.75*
- Subjectivity Lexicon (Wilson, et al., 2005)
  - Universal Subjective Clues
  - Consists of manually identified 8000 polar words with their prior polarity

Intermediary Tagged System (ITS)



A Meta-classifier is group of classifiers whose label predictions probabilities become the features for creating meta-classification model

H-I system architecture



Training Domain

Target Domain

Intermediary Tagger System (ITS)

Not all instances of Intermediary Tagged data are correct!

Intermediary Target Domain Data

Selection of
Optimal Intermediary Tagged Data

Labeled Target Data

High Accuracy Classifier (HAC)

Tagged Target Domain

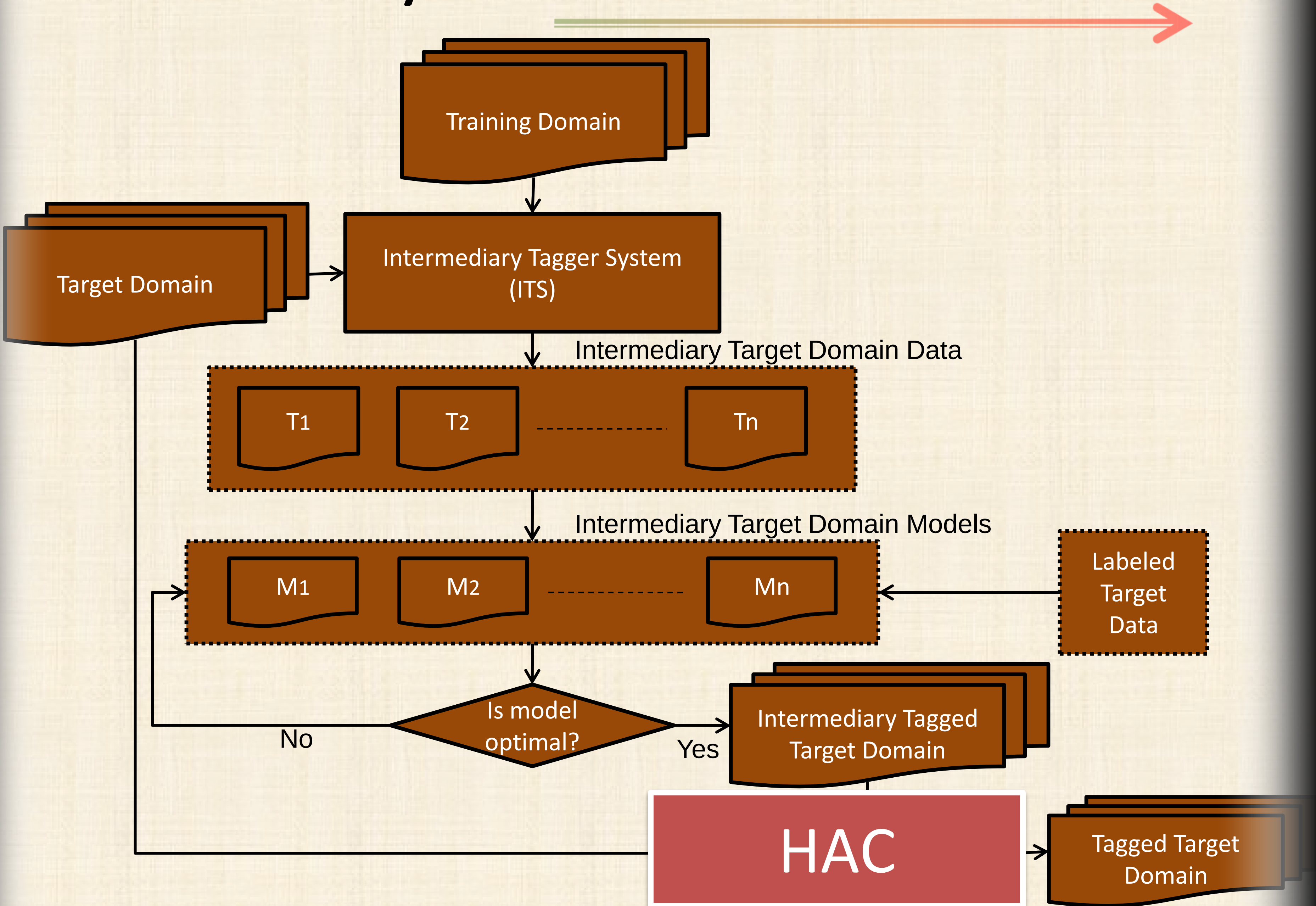
Selection of optimal intermediary tagged data

- Different sets of intermediary tagged target data created based on prediction probability generated by ITS
- Different SVM based models created using each set of tagged corpus
- Model tested on few hand labeled target data to find the best probability difference threshold
- This intermediary target set categorized as optimal intermediary tagged data

|Pr(+) – Pr(-)| > threshold ; 0.2<=threshold<1

Pr(.) = probability of document being positive or negative

H-I system architecture


Training Domain
Target Domain
Intermediary Tagger System (ITS)
Intermediary Target Domain Data
T1
T2
Tn
Intermediary Target Domain Models
M1
M2
Mn
Labeled Target Data
Is model optimal?
No
Yes
Intermediary Tagged Target Domain
HAC
Tagged Target Domain

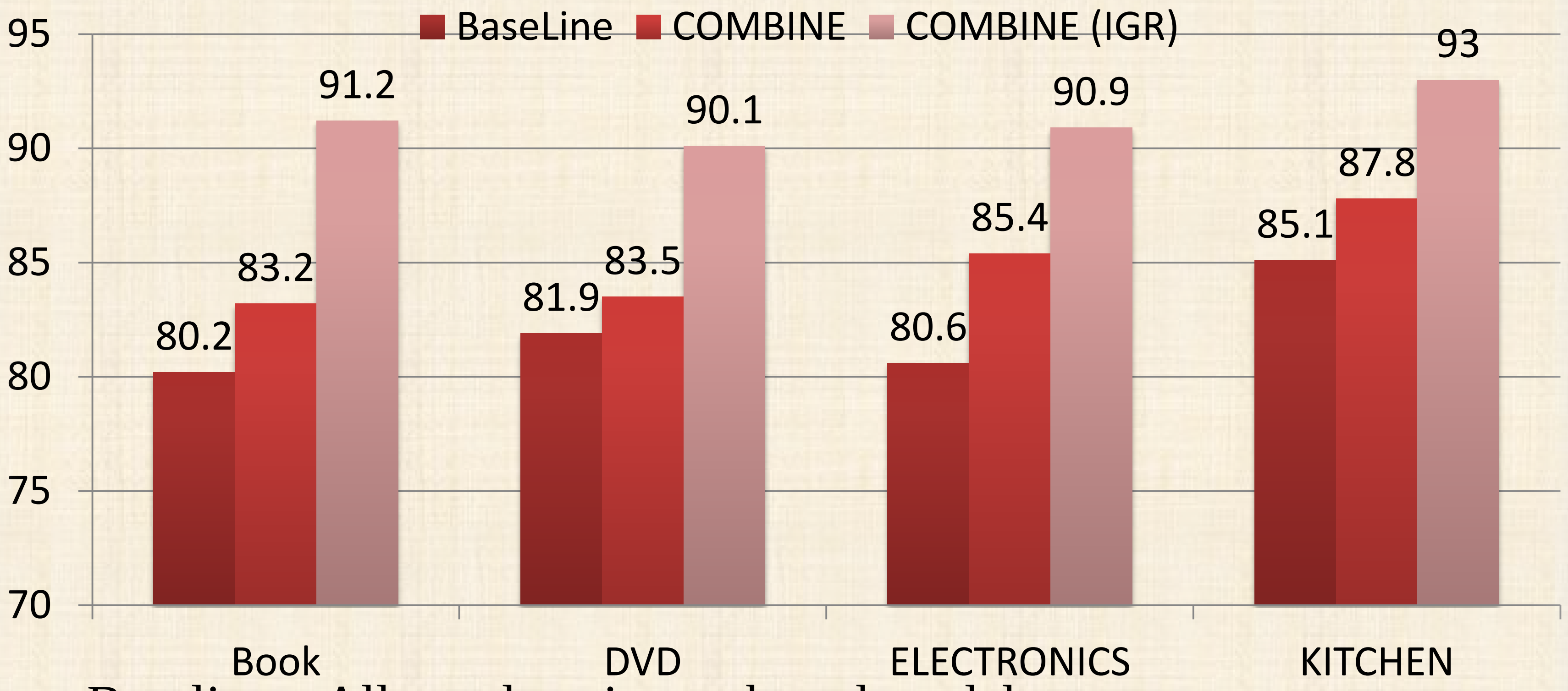
High Accuracy Classifier (HAC)

- Complex features not necessary for creating a High Accuracy Sentiment Classifier

- Intuition –
  - In a particular domain, there exist domain specific features (words/phrases) which have high discriminatory power
  - People use almost similar vocabulary in expressing their sentiment (pertinent to a domain)
- Process –
  - An Information Gain Ratio based feature selection done to select

Lame movie, Defective kitchen set

Results - HAC



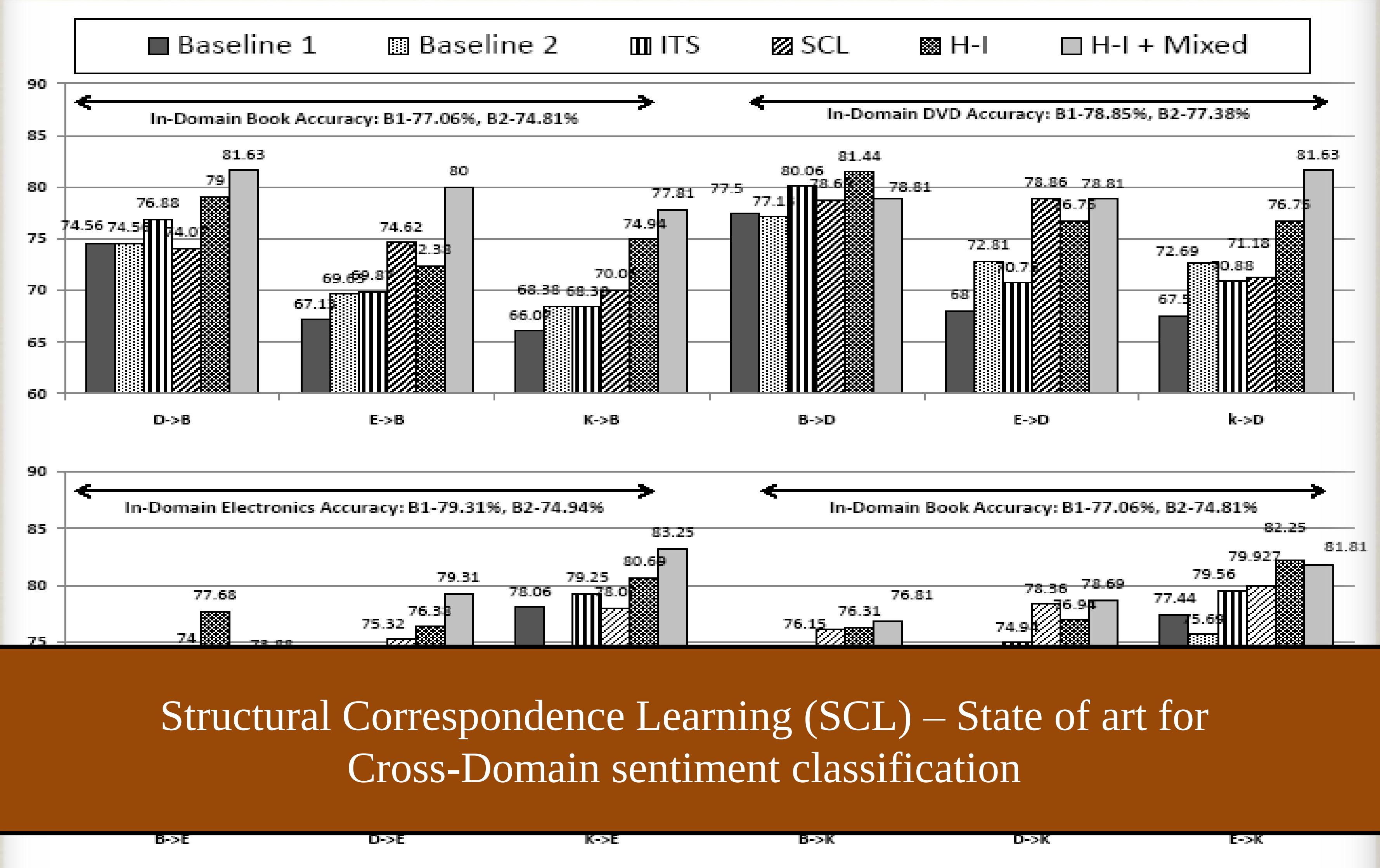
- Baseline – All words unigram-based model
- Combine – uni+bi+trigram model
- Combine (IGR) - uni+bi+trigram model after IGR
- An average increase of 10% in the classification accuracy with respect to the Baseline

Top features

DOMAIN	TOP IGR BASED FEATURES
BOOK (**B**)	waste of, love this, boring, stupid, too many, whatever, ridiculous, two stars
DVD (**D**)	worst, horrible, your money, lame, of the best, sucks, barely, ridiculous, save your, is a great, pathetic, dumb, not worth, ruined
ELECTRONICS (**E**)	return, terrible, waste your, highly, to return, poor, it back, returning, does not work, do not buy
KITCHEN (**K**)	easy to, easy to use. easy to clean, returning, waste your, tried to excellent, defective, horrible, poor, i love it

- Top IGR based features are different in different domains (Blitzer, et al., 2007).
- Domain specific nature of Sentiment Analysis

Results - Comparison



Structural Correspondence Learning (SCL) – State of art for Cross-Domain sentiment classification

Similarity between domains

- Cosine Similarity

	BOOK	DVD	ELECTRONICS	KITCHEN
BOOK	1	0.54	0.41	0.4
DVD	0.54	1	0.41	0.41
ELECTRONICS	0.41	0.41	1	0.49
KITCHEN	0.4	0.4	0.49	1

- Similar domains have high cross classification accuracy
- Dissimilar domains have relatively high increase in accuracy with respect to ITS performance

Conclusion

- Methodology for cross domain sentiment classification introduced
- An average cross-domain sentiment classification accuracy of 80% achieved
- Our system gives a high cross-domain Sentiment classification accuracy with an average improvement of 4.39% over the best baseline accuracy

Road map

New Domain

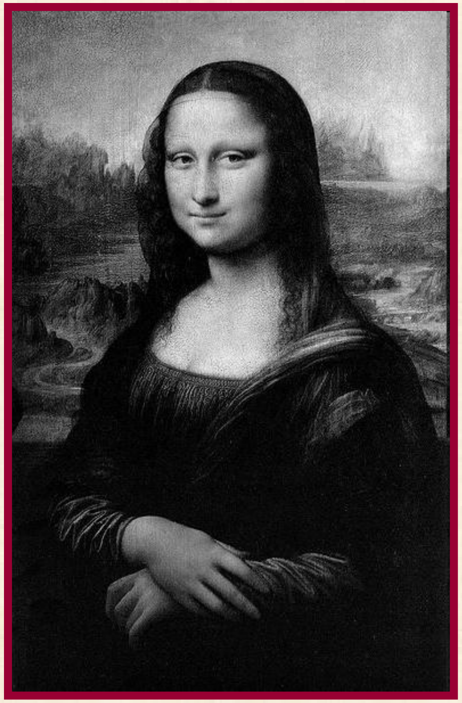
- Introduction
- Problem Statement
- Intuition
- Approach
- System Architecture
- Results

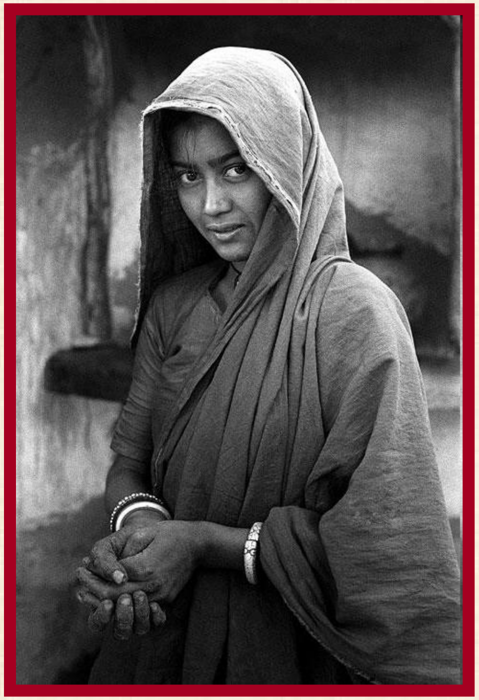
New Language

- Contribution
- Fall-back strategy

New Text Form

Sentiment Analysis for a new language





Based on our work titled ‘A Fall-back Strategy for Sentiment Analysis in a New Language: a Case Study for Hindi’ to be presented at International Conference on Natural Language Processing (ICON) 2010 to be held in December 2010

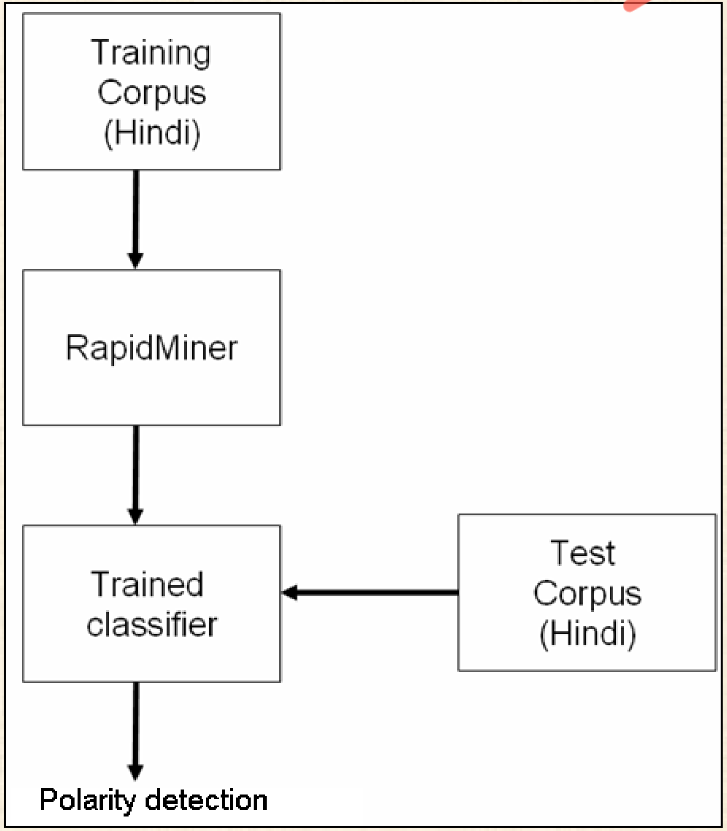
Contribution

A **first** study/implementation of Sentiment analysis for Hindi

1. Creation of a manually annotated corpus for Hindi
2. Creation of **Hindi SentiWordNet**, based on the equivalent for English
3. **Fall-back strategy** for adapting sentiment analysis to a new language

In-language SA

- Train on Hindi
- Test on Hindi
- Classifier: SVM



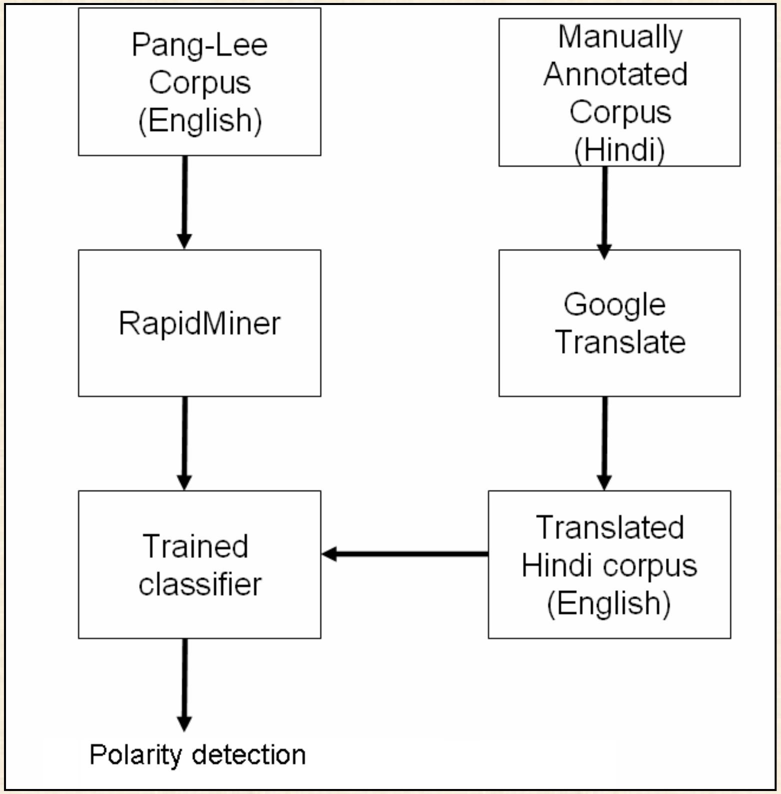
- Does ‘rich cousin’ English help the task more than the scarcely-resourced Hindi?

MT-based SA

- ‘Good’ translation not expected as long as important sentiment-bearing words get translated correctly
- Classifier: SVM

Issue:
वह अनंत के पिछले नायक इमरान की तरह दबंग नहीं दिखते
(He does not look tough like Anant’s last hero Imran)

‘It's infinite, like the last hero Imran not look sturdy’.


Pang-Lee Corpus (English)
Manually Annotated Corpus (Hindi)
RapidMiner
Google Translate
Trained classifier
Translated Hindi corpus (English)
Polarity detection

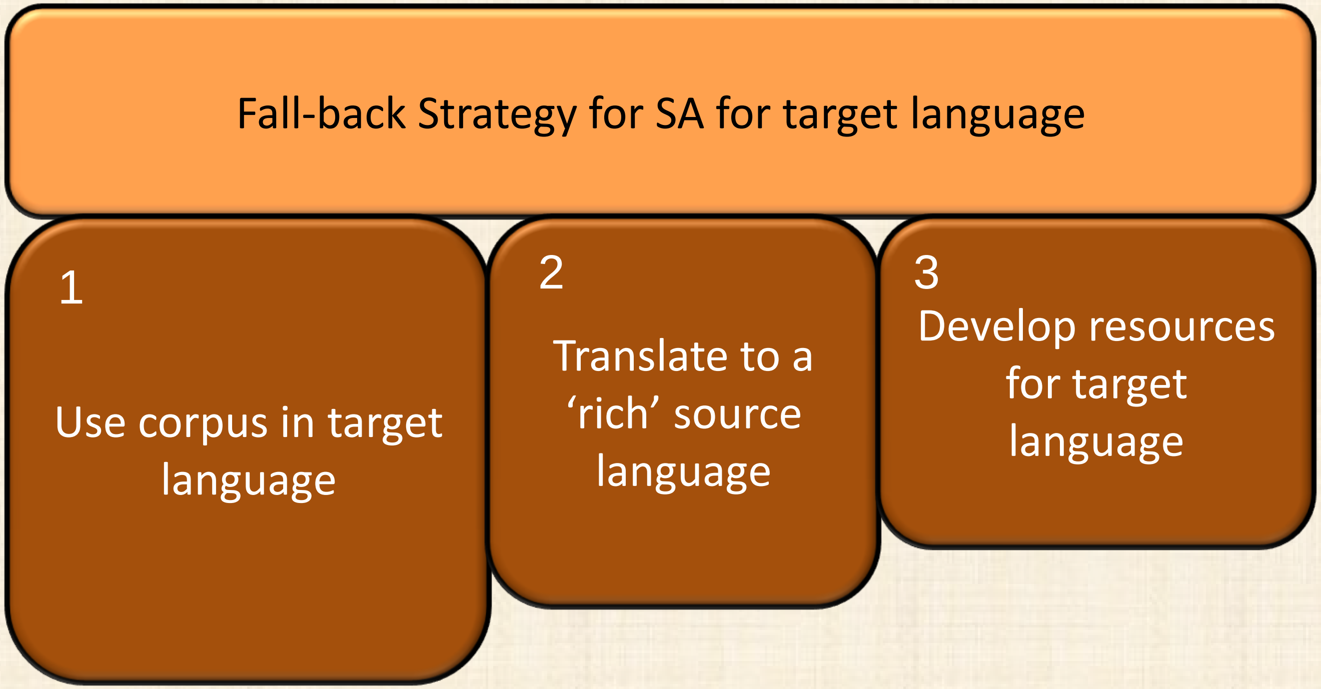
Resource-based SA

Using H-SWN for SA

- For each word in the document,
  1. Apply stop word removal and stemming
  2. Look up the sentiment triple for each word in the H-SWN.
  3. Assign to a word the polarity whose score is the highest.
- Assign to a document the polarity which majority of its words possess.

Fall-back strategy

Alternatives	Accuracy (%)
In-language SA	78.14
MT-based SA	65.96
Resource-based SA	60.31



Fall-back Strategy for SA for target language

1 Use corpus in target language

2 Translate to a ‘rich’ source language

3 Develop resources for target language

Road map

New Domain

- Introduction
- Problem Statement
- Intuition
- Approach
- System Architecture
- Results

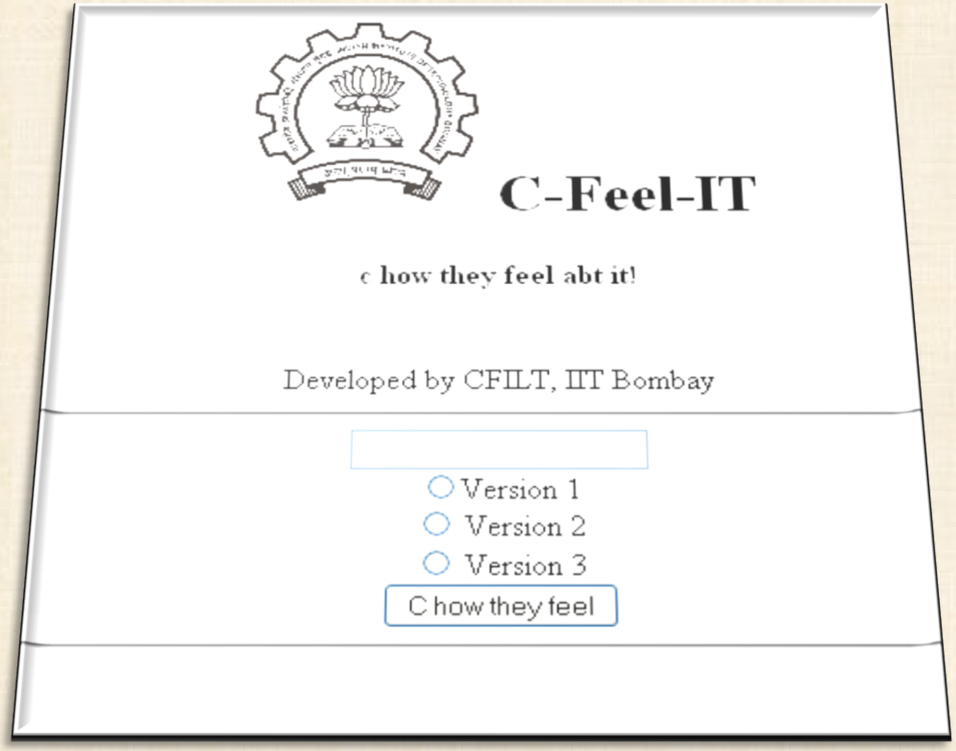
New Language

- Contribution
- Fall-back strategy

New Text Form

- Features of tweets
- Working
- Three versions of C-FeeL-IT
- Experiments & Heuristics

Sentiment Analysis for a new text form



C-Feel-IT

c how they feel abt it!

Developed by CFILT, IIT Bombay

○ Version 1
○ Version 2
○ Version 3

Our implementation of the problem suggested by Dr. Rajat Mohanty

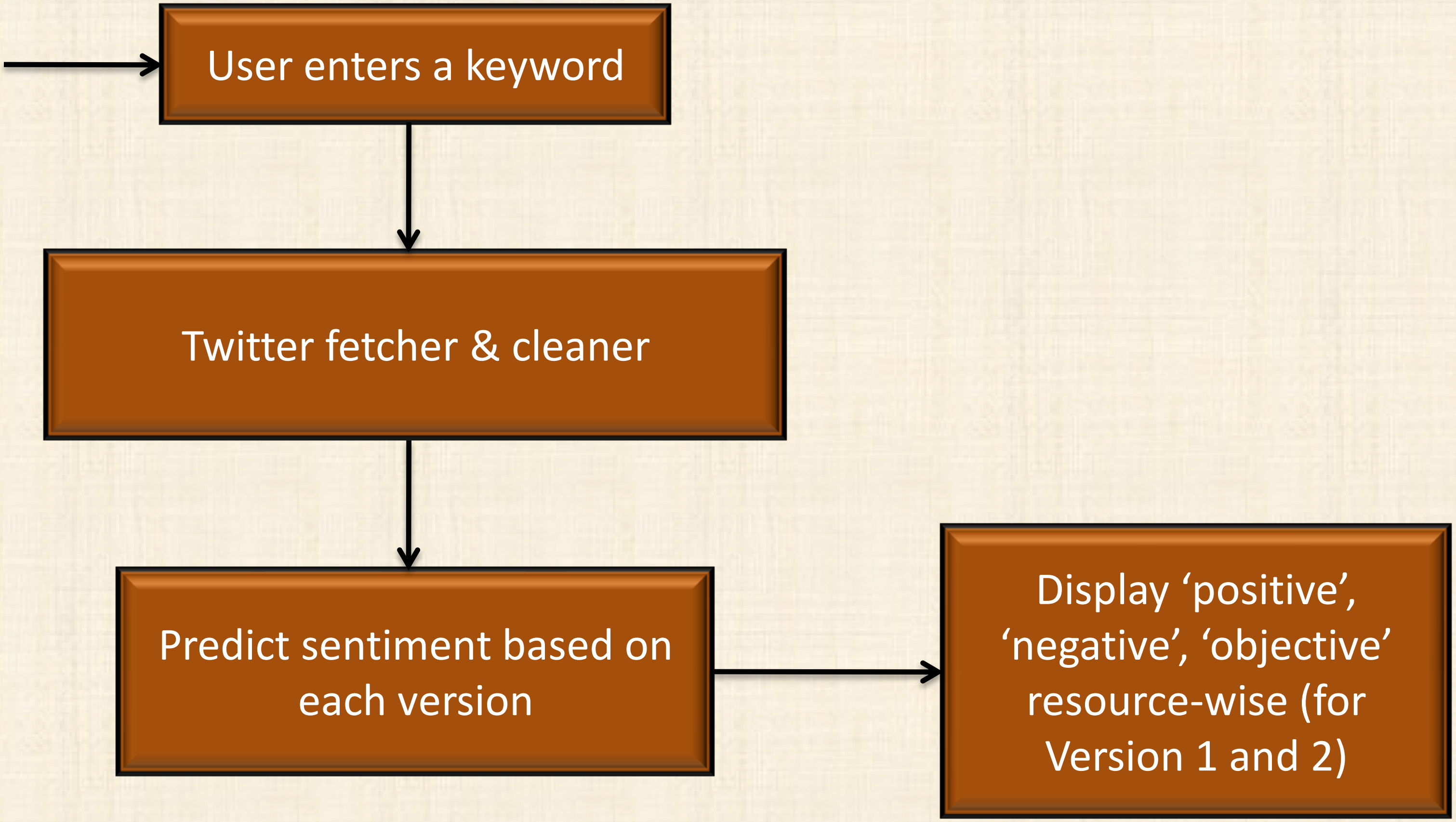
Available at: http://www.clia.iitb.ac.in/TwitterApp/index.jsp

New Text form: Tweets

Tweets as opposed to **blog posts/reviews**:

- Short
- Links, smileys
- Extensions of words (‘haapppyy’ for ‘happy’)
- Contractions of words (‘abt’ for ‘about’)
- Difficult for construction of a feature vector to use a machine learner

Working of C-FeeL-IT



User enters a keyword

Twitter fetcher & cleaner

Predict sentiment based on each version

Display ‘positive’, ‘negative’, ‘objective’ resource-wise (for Version 1 and 2)

Resources used

- SentiWordNet (Andrea & Sebastani,2006)
- Subjectivity clues (Weibi et al, 2004)
- Taboada (Taboada & Grieve, 2004)
- Inquirer (Stone et al, 1966)

Versions

- **Version 1:** ‘Individual words of a tweet constitute the sentiment of a tweet.’

- **Version 2:** ‘Certain POS bi-tags indicate sentiment more than others.’

- **Version 3:** ‘A regression classifier can help predict the sentiment of a document constructed by stitching together tweets.’

How do we combine results of resources?

- Accuracy of each of the resources calculated on RTPolarity (Pang and Lee, 2005) dataset
- Weight each of them for each label as:

Weight of a resource for label X= No. of correct instances for label X / Total no. of instances of label X

- Normalized for the three labels

Combining results of resources

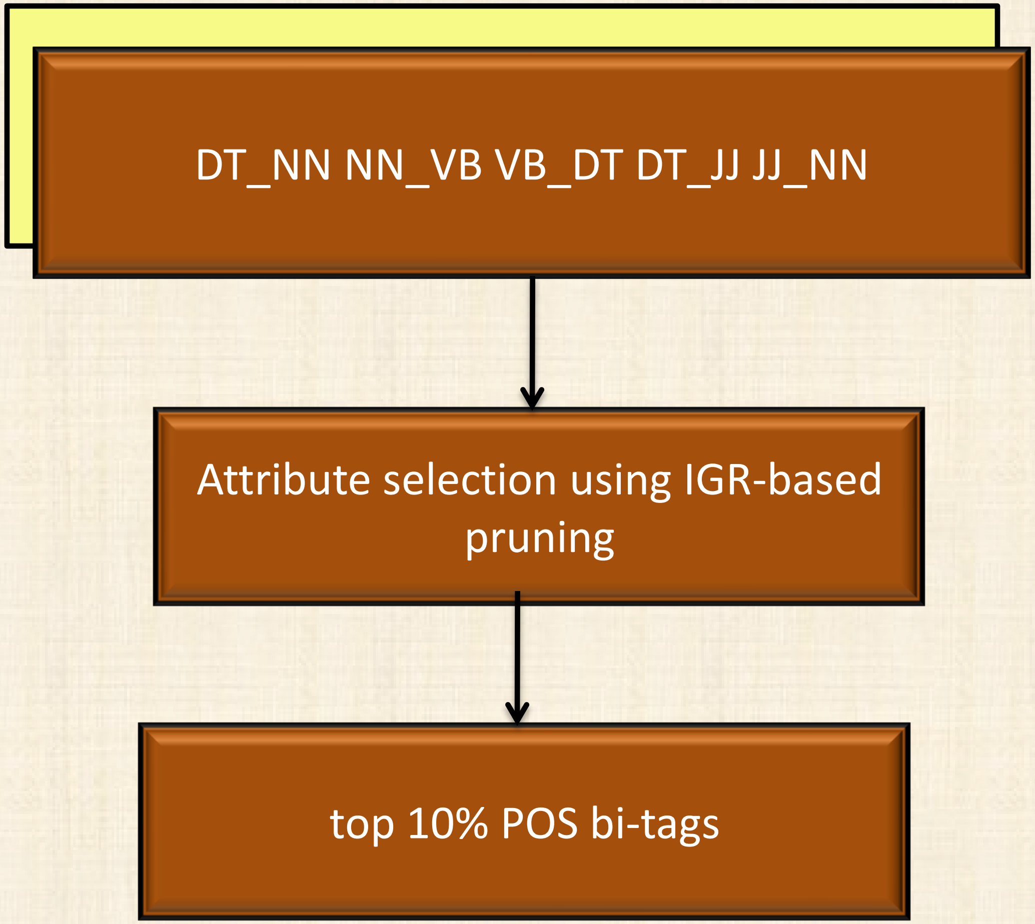
Weights of resources

Sentiment Label	SentiWordNet	Pittsburgh	Inquirer	Taboada
Positive	0.71	0.739	0.786	0.996
Negative	0.21	0.49	0.23	0.013

For objective = 1

Predictions are normalized at the end to a consolidated prediction

How do we select POS tags?



DT_NN NN_VB VB_DT DT_JJ JJ_NN

Attribute selection using IGR-based pruning

top 10% POS bi-tags

Heuristics

- Smileys pinpoint sentiment
- Negations invert sentiment till a window
- Extensions indicate magnitude of sentiment

Demonstration

Conclusion

Looking at three new turfs in which an existing SA system can be extended to:

- **A new domain**: Novel technique for cross-domain sentiment prediction
- **A new language**: Fall-back strategy for sentiment analysis in a new language (case taken: Hindi)
- **A new text form**: Entity-based search engine for analysing tweets for sentiment

Road ahead

New Domain: Better intermediary tagged system to gain better classification

New Language: Better assistance from resource-rich language by transitioning to concept space

New Text form: Better regression model to achieve close-to-reality prediction

Reference (1/2)

- Anthony Aue and Michael Gamon,2005,Customizing Sentiment Classifiers to New Domains: A Case Study, Proceedings of Recent Advances in NLP
- Alina Andreevskaia and Sabine Bergler,2008,When Specialists and Generalists Work Together: Overcoming Domain Dependence in Sentiment Tagging, Proceedings of ACL-08: HLT,PP 290-298,Columbus, Ohio
- Blitzer, John and Dredze, Mark and Pereira, Fernando, Biographies, Bollywood, Boom-boxes and Blenders: Domain Adaptation for Sentiment Classification,Proceedings of the 45th Annual Meeting of ACL,2007,pages = 440--447,Prague, Czech Republic
- Andrea Esuli and Fabrizio Sebastiani,SentiWordNet: A Publicly Available Lexical Resource for Opinion Mining,Proceedings of LREC-06,2006,Genova, Italy
- Theresa Wilson and Janyce Wiebe and Paul Hoffmann,Recognizing Contextual Polarity in Phrase-Level Sentiment Analysis,Proceedings of EMNLP,2005,pages =347-354,Vancouver, Canada
- Bo Pang and Lillian Lee and Shivakumar Vaithyanathan,Thumbs up? Sentiment Classification Using Machine Learning Techniques,Proceedings of EMNLP,2002,pages = 79-86,Philadelphia, Pennsylvania
- David H. Wolpert,Stacked generalization,Neural Networks,Volume 5, 1992, Pages=241-259
- Chih-Chung Chang and Chih-Jen Lin,LIBSVM: a library for support vector machines,2001
- Wu, Ting-Fan and Lin, Chih-Jen and Weng, Ruby C.,Probability Estimates for Multi-class Classification by Pairwise Coupling,Journal of Machine Learning and research,Vol 5,2004},975--1005

Reference (2/2)

- Bo Pang and Lillian Lee and Shivakumar Vaithyanathan,Thumbs up? Sentiment Classification Using Machine Learning Techniques,Proceedings of EMNLP,2002,pages = 79-86,Philadelphia, Pennsylvania

- David H. Wolpert,Stacked generalization,Neural Networks,Volume 5, 1992, Pages=241-259

- Chih-Chung Chang and Chih-Jen Lin,LIBSVM: a library for support vector machines,2001

- Wu, Ting-Fan and Lin, Chih-Jen and Weng, Ruby C.,Probability Estimates for Multi-class Classification by Pairwise Coupling,Journal of Machine Learning and research,Vol 5,2004},975—1005

- Read, Jonathon,Using emoticons to reduce dependency in machine learning techniques for sentiment classification, ACL '05: Proceedings of the ACL Student Research Workshop},2005,pages =43--48,Annrbor, Michigan,Morristown, NJ, USA
- Bo Pang and Lillian Lee, 2005, Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales, Proceedings of the ACL.
- Wiebe, J., Wilson, T., Bruce, R., Bell, M. & Martin, M. ,Learning Subjective Language, 2004,Computational Linguistics .
- Philip J. Stone, Dexter C. Dunphy, Marshall S. Smith, and Daniel M. Ogilvie. The General Inquirer: A Computer Approach to Content Analysis. MIT Press, 1966.

- Taboada, M. and J. Grieve (2004) Analyzing Appraisal Automatically. American Association for Artificial Intelligence Spring Symposium on Exploring Attitude and Affect in Text. Stanford. March 2004. AAAI Technical Report SS-04-07. (pp.158-161).