Learning More with Less

Reducing Annotation Effort with Active and Interactive Learning

Shilpa Arora & Manas Pathak
Advisor: Dr. Eric Nyberg

Student Research Symposium
September 14th, 2007

Language Technologies Institute
Carnegie Mellon University

Outline

- Introduction
- Motivation
- Objectives
- Approach
- Evaluation Measures
- Experiments & Results
- Conclusion
- Future Work

Introduction

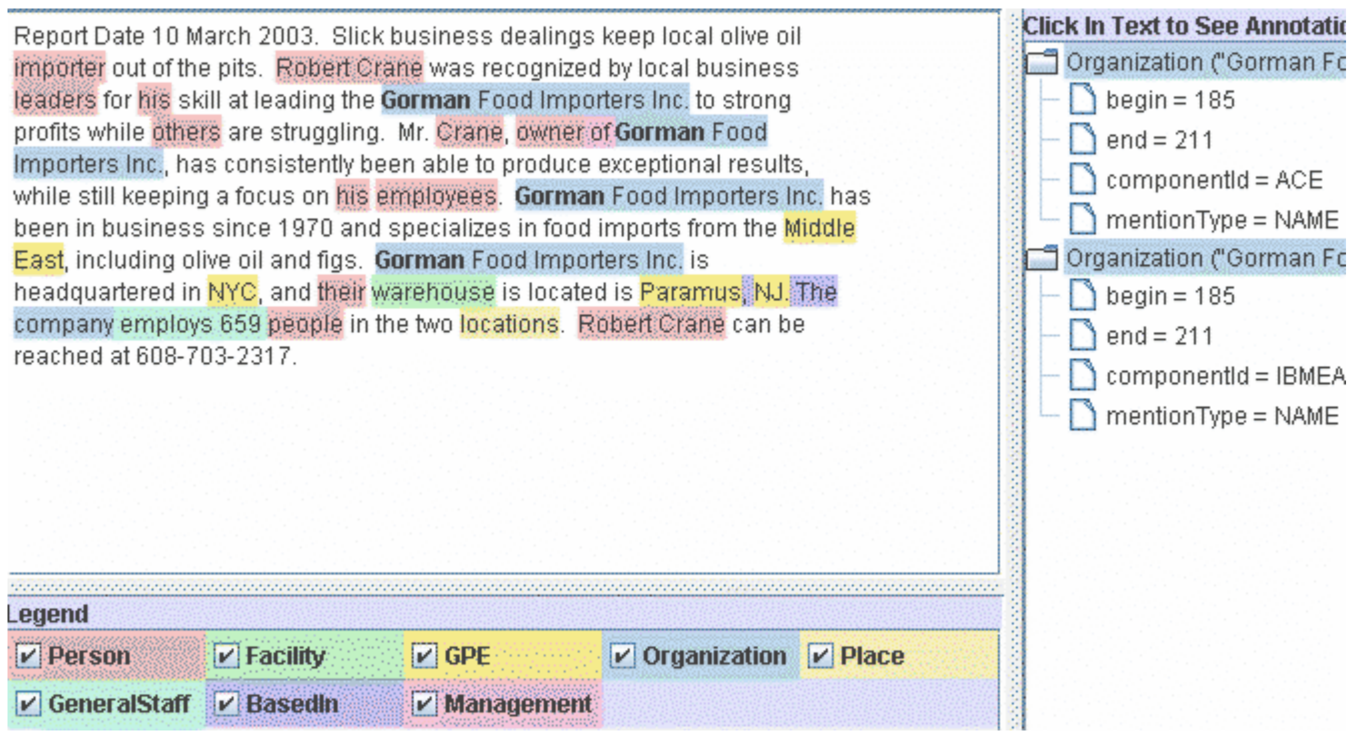
- Text Annotation a.k.a Information Extraction

Examples:

Simple/binary: Classification (Spam or not)

Multi-class: Named Entity Recognition (NER), Part of Speech (POS) Tagging

Complex/Structured: Semantic Role Labeling, Event Extraction



How are *Annotations* learnt?

- Hand-coded Rules
  - Need lots of rules, domain experts & doesn't generalize well

- Statistical Machine Learning Approach
  - Requires a lot of pre-annotated training data
  - Annotating text is a *time consuming*, *tedious*, *error prone* process
  - All examples are not equally *informative* or equally *easy* to annotate

Ben's boss has asked him to annotate corpus with company establishment events

Ben

Ben's boss has asked him to annotate corpus with company establishment events

Ben

Traditional batch annotation process

• Human annotators must exhaustively and completely annotate large amounts of data

• Requires a lot of user's effort

• Much of this effort could be unnecessary if it doesn't help the learner.

Ben's boss has asked him to annotate corpus with company establishment events

Ben

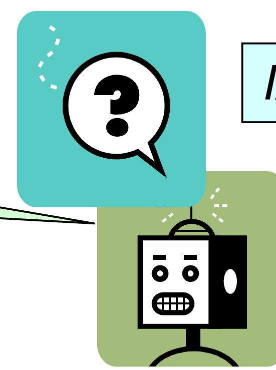
Traditional batch annotation process

- Human annotators must exhaustively and completely annotate large amounts of data
- Requires a lot of user's effort
- Much of this effort could be unnecessary if it doesn't help the learner.

Interactive annotation process

- Learner suggest annotations it already thinks/knows should be labeled
- User confirms and corrects automatic annotations
- Learner recommends documents that will help it learn => *Learner can ask questions*
- User can see if and how their effort is being utilized

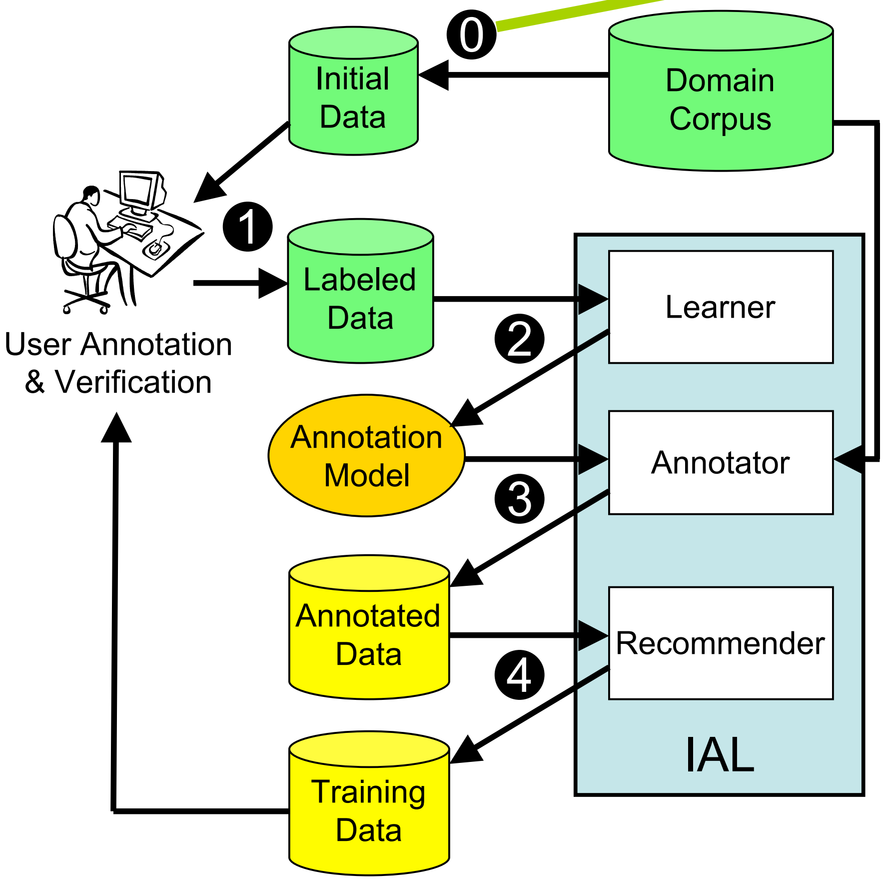
Ben! I am ***confused*** about a few examples, can you help? "*Microsoft established a set of certification programs …*"- does this event also talk about company establishment?



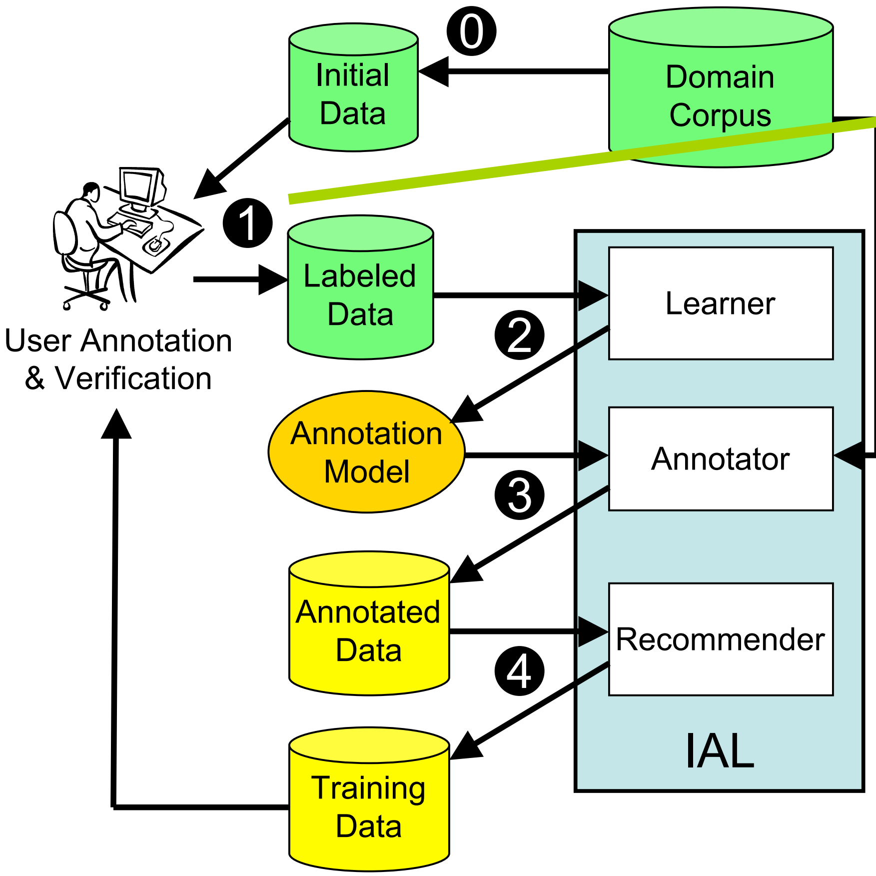
IAL

Interactive Annotation Learning

Ben selects an initial example set

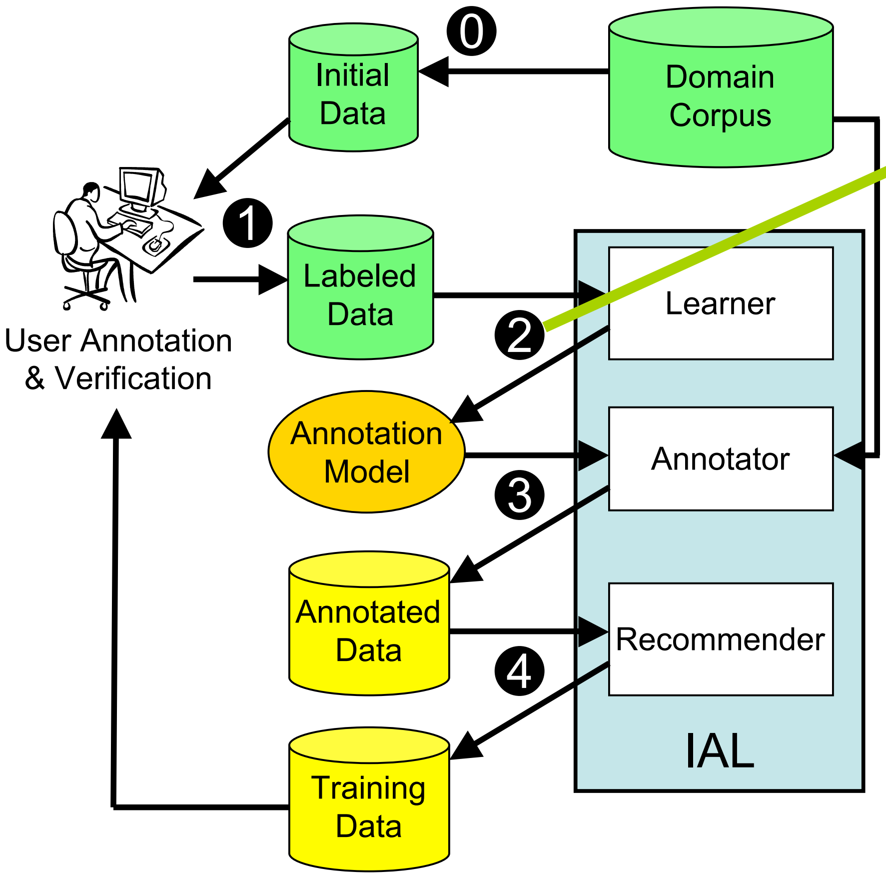


Interactive Annotation Learning



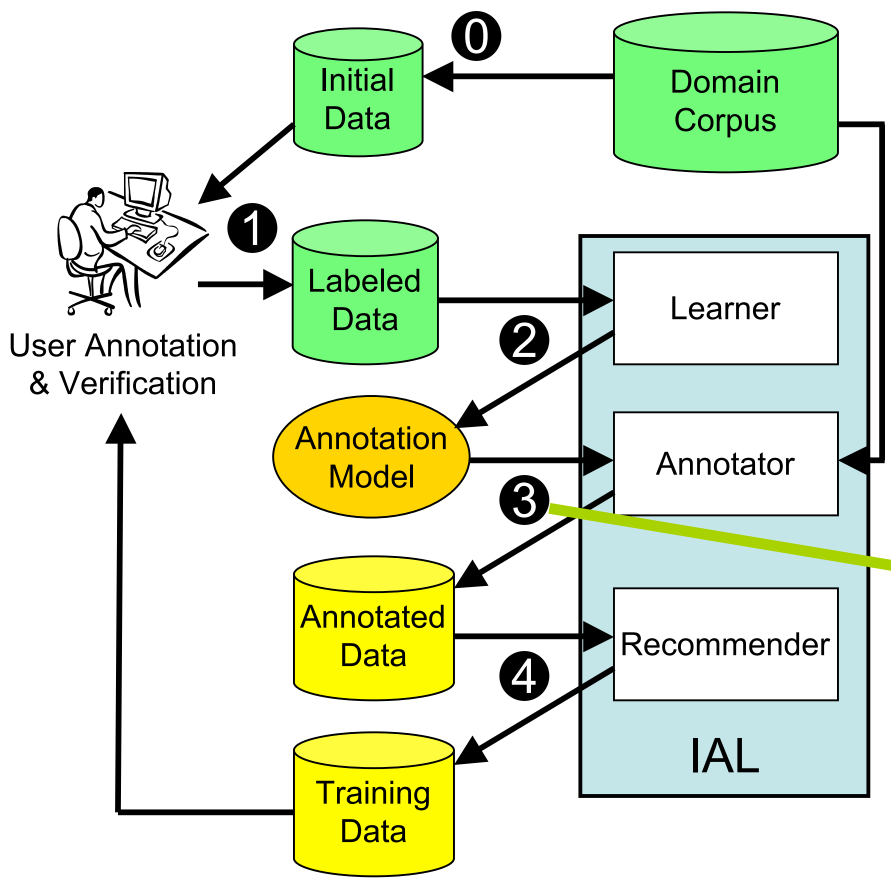
He annotates the structures he wants the system to learn

Interactive Annotation Learning



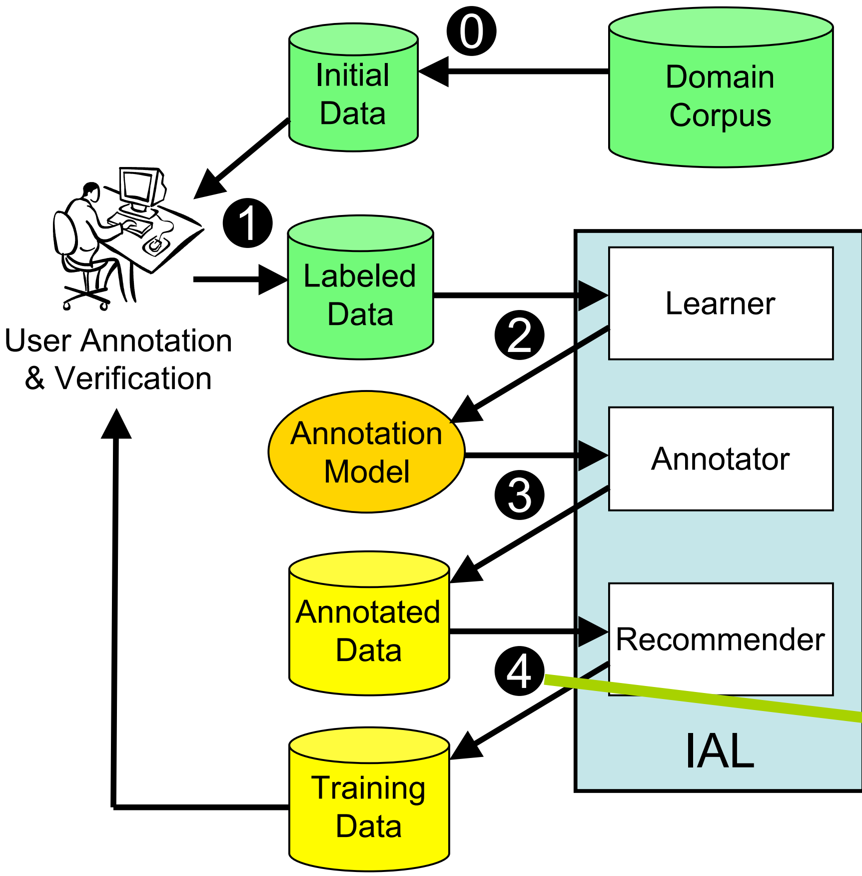
The labeled data are input to a statistical learner that creates an annotation model

Interactive Annotation Learning



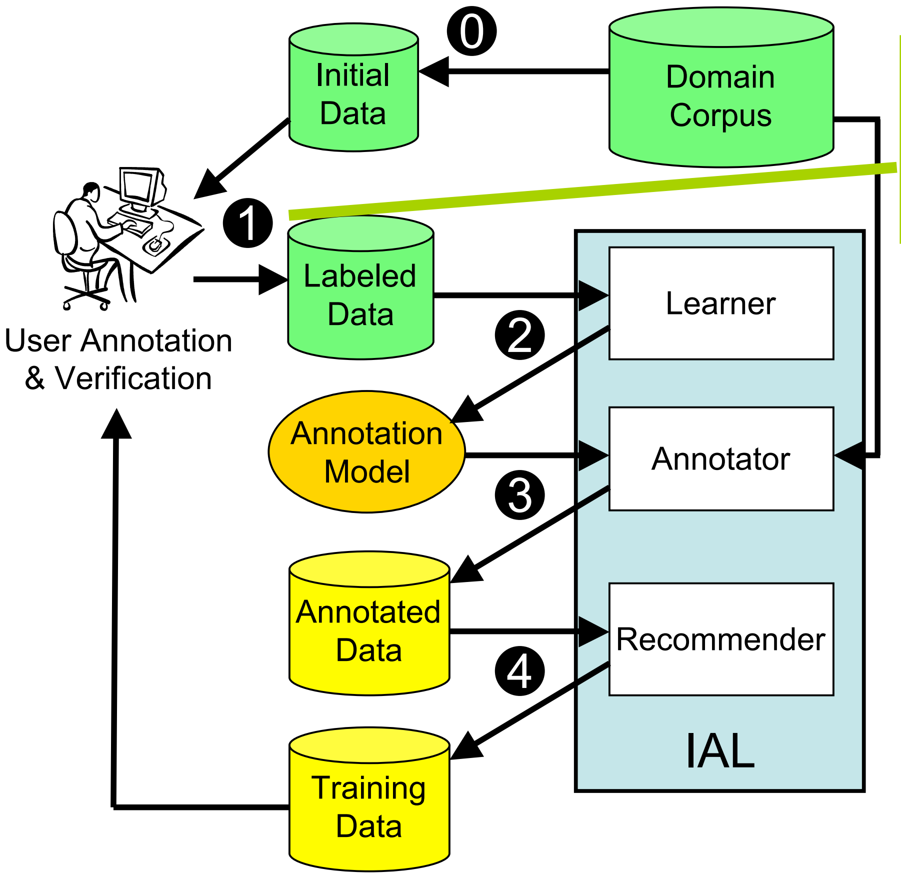
The learned annotator is used to annotate an additional set of documents

Interactive Annotation Learning



System selects the most appropriate subset of the annotated examples for Ben's verification

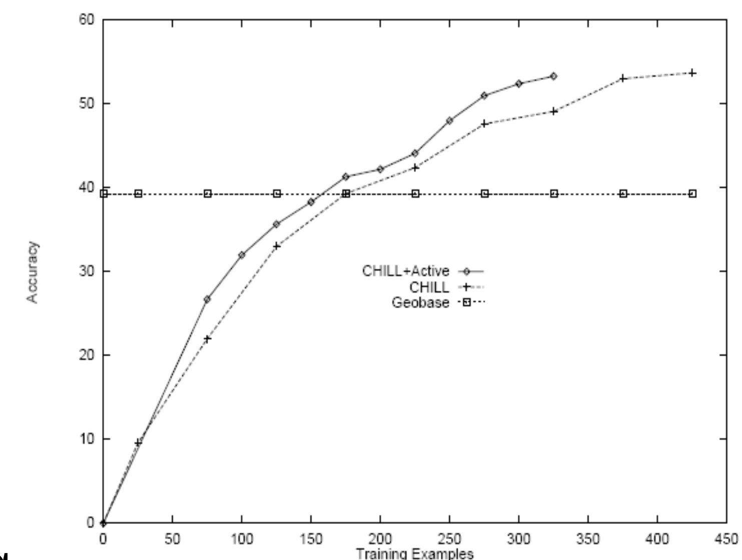
Interactive Annotation Learning



Ben verifies or corrects the automatic annotations, deletes the wrong ones and adds the missing ones

Missing Science

- Learning should happen naturally (= “in task”)
- Interactive Learning
  – User in the loop learning
  – User sees the result of his effort
- Active Learning
  – Faster convergence
- Interactive Learning
  – Minimize user effort
- ***Interactive Active Learning***
  – Best of both the worlds !
  – User effort as an Evaluation measure [Kristjannson et.al.] & ***Recommendation Strategy (New!)***

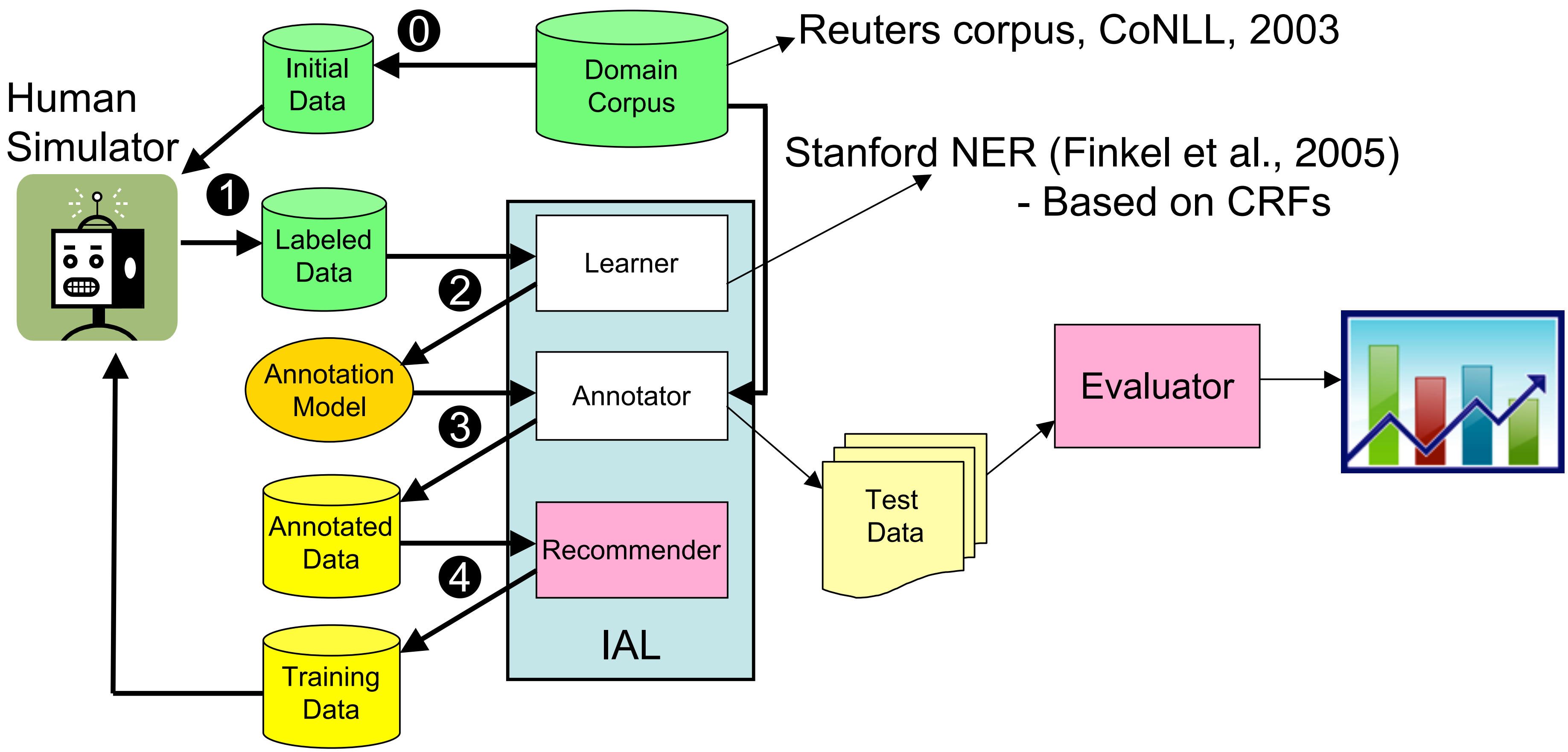


Objective

- Hypothesis
  - *"There exists a combination of Active & Interactive learning recommendation strategies that performs significantly better than random selection in both accuracy and user-effort measures."*

- Prove that the hypothesis holds for an example problem: Named Entity Recognition
  - Recommendation strategies
  - Combination of several strategies
  - Evaluation measures

Interactive Annotation Learning Framework



Initial Software Framework: (Ben Lambert & Jose Alavedra, SE II project)

Approach

- Human Simulator
  – Gold Standard (Perfect or Imperfect?)
- Recommenders: Selective Sampling Strategies:
  - **Uncertainty based** *[Thompson et. al., Culotta et. al …]*
    – *That the model is most uncertain about*
  - **Committee based** *[McCallum et. al. …]*
    – *With most disagreement among committee of classifiers*
  - **Diversity based** *[Shen et. al., Seokhwan et. al….]*
    – *Different from those already in the training pool*
  - **Representative power based** *[Shen et. al.]*
    – *Most representative of other examples*
  - **User effort based**
    – *Easier to annotate*

Approach

- Human Simulator
  - Gold Standard (Perfect or Imperfect ?)
- Recommenders
  - Selective Sampling Strategies:
    - Uncertainty based *[Thompson et. al., Culotta et. al …]*
    - Committee Based Methods *[McCallum et. al. …]*
    - Diversity based *[Shen et. al., Seokhwan et. al….]*
    - Representative *[Shen et. al.]*
    - User effort based

Uncertainty based Recommenders

- Average Annotation Confidence (AC)

$$AC = \frac{\sum_{i=1}^{n} conf(l_i)}{N}$$

where: $conf(l_i)$ = confidence assigned to annotation l_i

N = number of annotations

- Relative number of Annotations below threshold

$$RBT = \frac{t - min_t}{max_t - min_t}$$

$$\text{and, threshold } th = \frac{\sum_{i=1}^{D} AC_i\, n_i}{\sum_{i=1}^{D} n_i}$$

where: D = number of Documents

n_i = number of annotations in document i

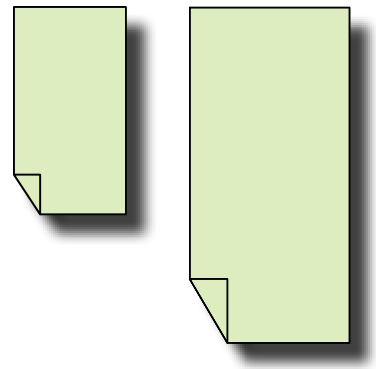
t = number of annotations with confidence below threshold th

User Effort Based Recommenders

- Relative Document Length

$$RDL = \frac{d - min_d}{max_d - min_d}$$

where: d = number of words



- Annotations Density

$$AD = \frac{\#\text{words in annotations}}{\#\text{words in document}}$$

Composite Recommender

- Combining several recommenders where each addresses different concerns
- How do we combine the result of two recommenders
  - **Weighted sum of scores** [Shen et. al.]

    $\alpha A + (1 - \alpha) B$
  - **Two-stage approach** [Shen et. al.]

    Output of one recommender => Input of another
  - **MMR Combination** [Seokhwan et. al.]

    $$score(s_i) \stackrel{def}{=} \lambda * \text{Uncertainty}(s_i, M) - (1 - \lambda) * \max_{s_j \in T_M} \text{Similarity}(s_i, s_j)$$
  - **Weighted Random Sampling** [Sarawagi et. al.]

    Weight each instance by its uncertainty & do weighted sampling - preserves underlying data distribution

Composite Recommender

- Combining several recommenders where each addresses different concerns
- How do we combine the result of two recommenders
  - Weighted sum of scores [Shen et. al.]
  - Two-stage approach [Shen et. al.]
  - MMR Combination [Seokhwan et. al.]
  - Weighted Random Sampling [Sarawagi et. al.]

Weighted Combination of Active & Interactive Strategies

Evaluation Measures

- Annotation F-measure

$$F = \frac{2 \times precision \times recall}{precision + recall}$$

- Expected Number of User Actions (ENUA)
  [Kristjannson et al., 2004]
  – An estimation of user effort
  – Calculated by the human annotator simulator by comparing the annotations made with Gold Standard

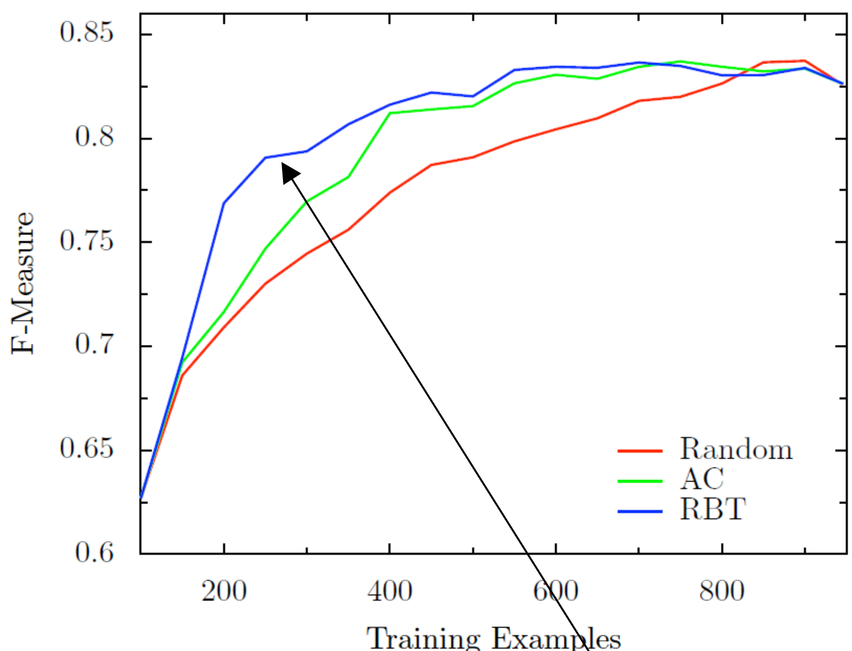
Experiments

- Data: Reuters Corpus w. Named Entities CoNLL, 2003 [Sang et al., 2003]
  - Training set: 900 documents
  - Test set: 245 documents

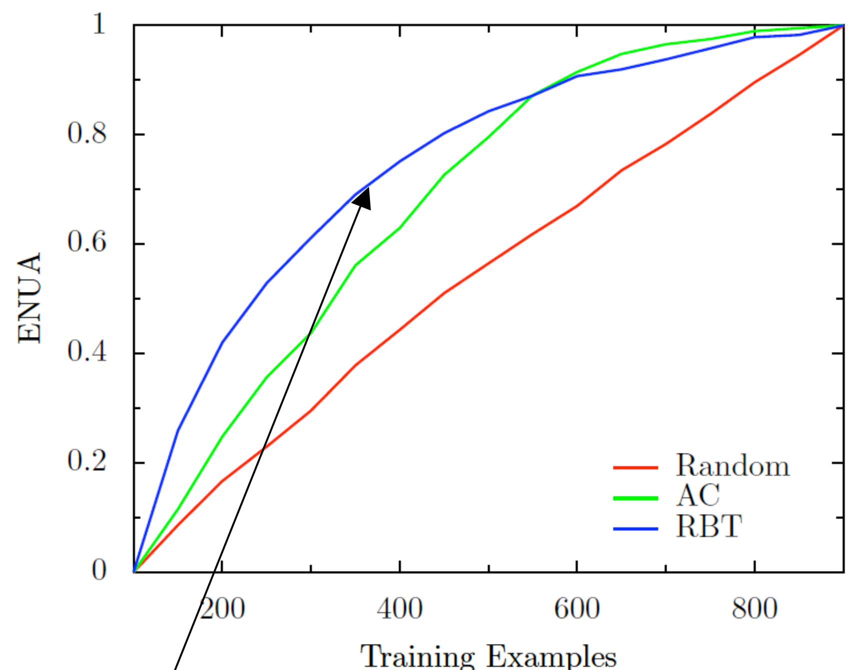
- Evaluation
  - Different recommendation strategies
  - Baseline: Random

Convergence Curves

AC (blue) - RBT (green) - Random (red)



F-Measure



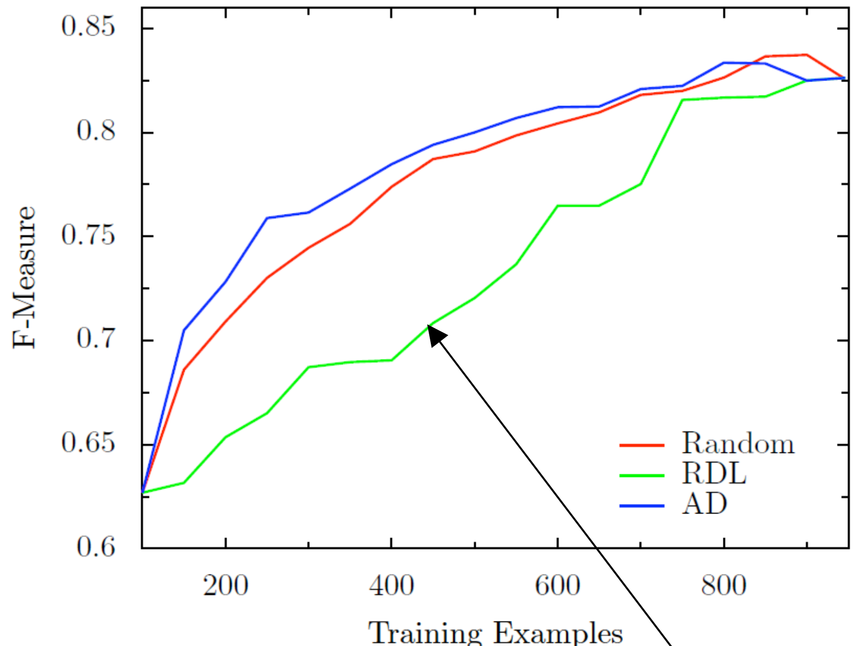
ENUA

- AC: Avg. annotation confidence
- RBT: Rel. docs below threshold
- RDL: Rel. doc length
- AD: Annotation density

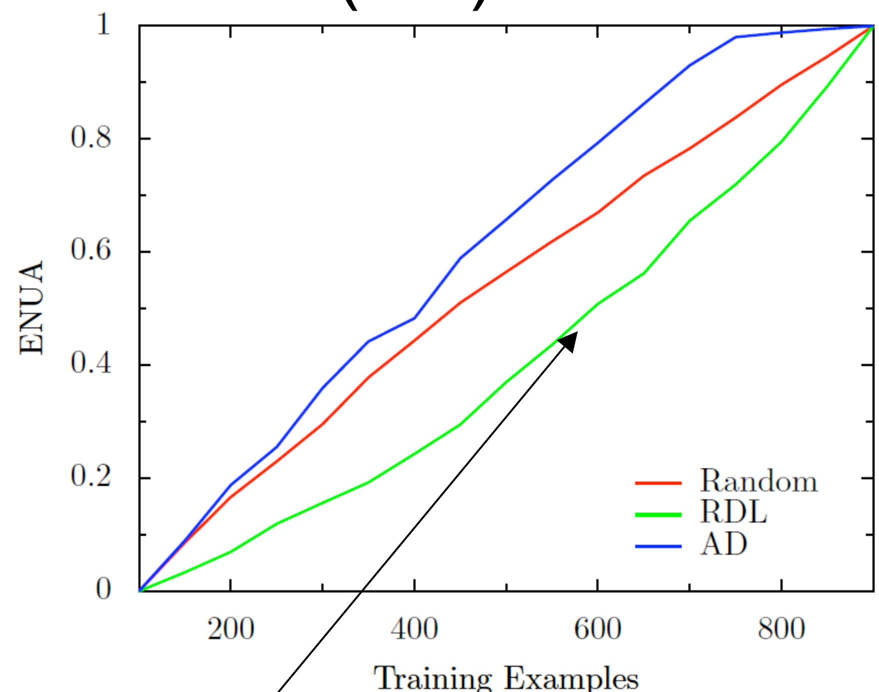
RBT does good in F-Measure (significance $p<0.05$) but poorly in ENUA

Convergence Curves

AD (blue) - RDL (green) - Random (red)



F-Measure



ENUA

- AC: Avg. annotation confidence
- RBT: Rel. docs below threshold
- RDL: Rel. doc length
- AD: Annotation density

RDL does good in ENUA (significance $p<0.05$) but poorly in F-measure

Convergence Curves

- Problem?

  No recommendation strategy performing well for both measures

- Solution:

  Weighted combination of different strategies

The Right Balance

- Experiment for all combinations of RDL & RBT

ΔENUA (blue) & ΔF-Measure (red)

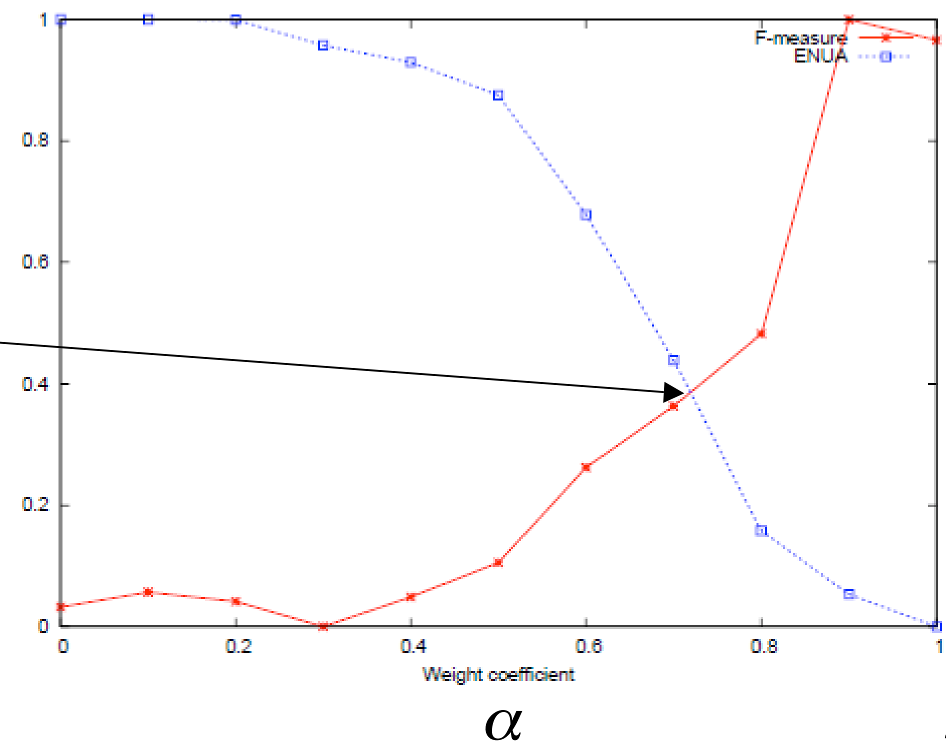
$$\alpha RDL + (1-\alpha) RBT$$

$$\Delta E_i = E_{max} - E_i$$

$$\Delta F_i = F_{max} - F_i$$

Cross-over point

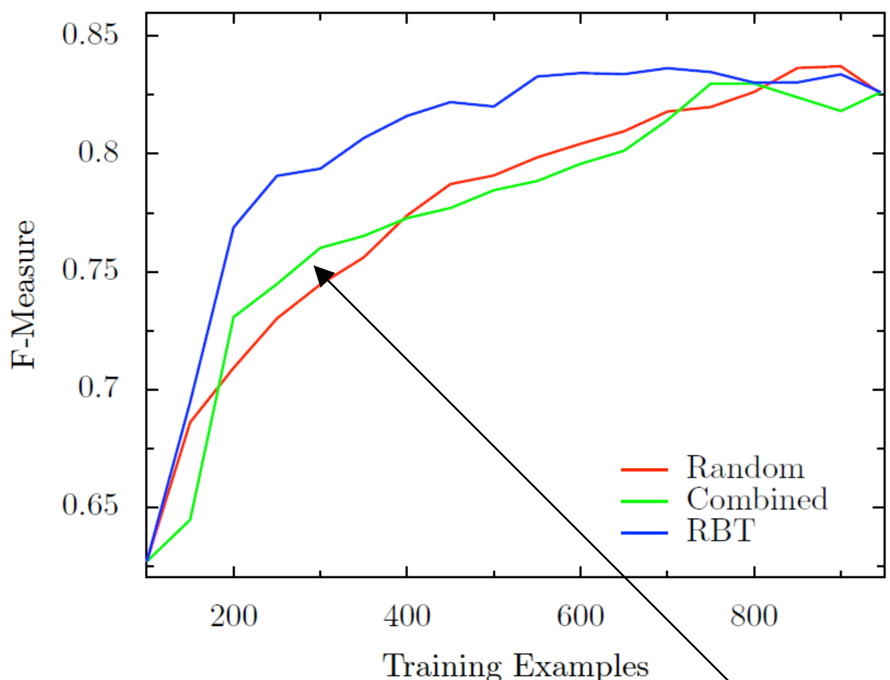
0.7 RDL + 0.3 RBT



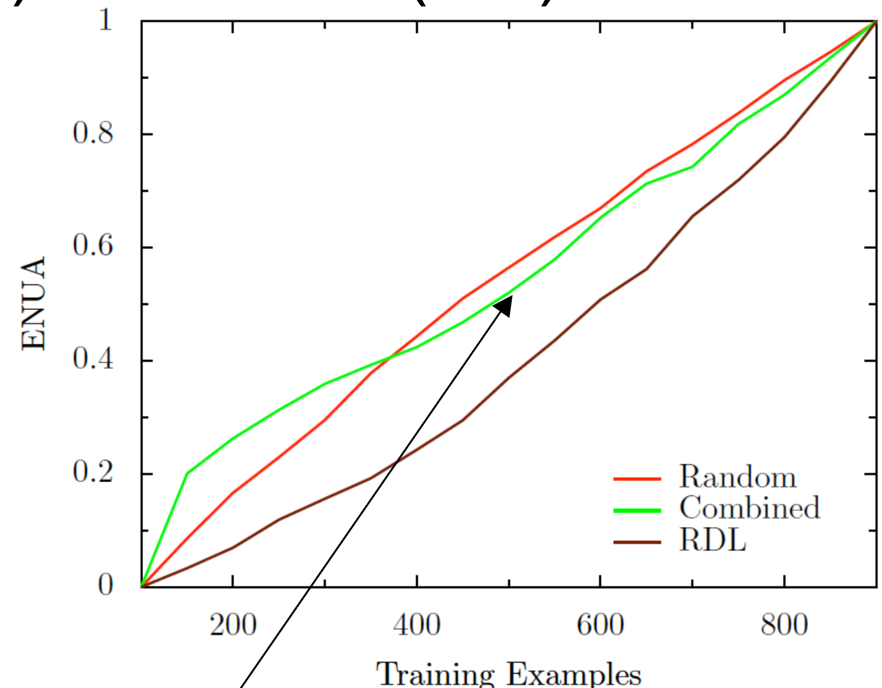
- AC: Avg. annotation confidence
- RBT: Rel. docs below threshold
- RDL: Rel. doc length
- AD: Annotation density

Combined Strategy

0.7 RDL + 0.3 RBT (green) - Random (red)



F-Measure



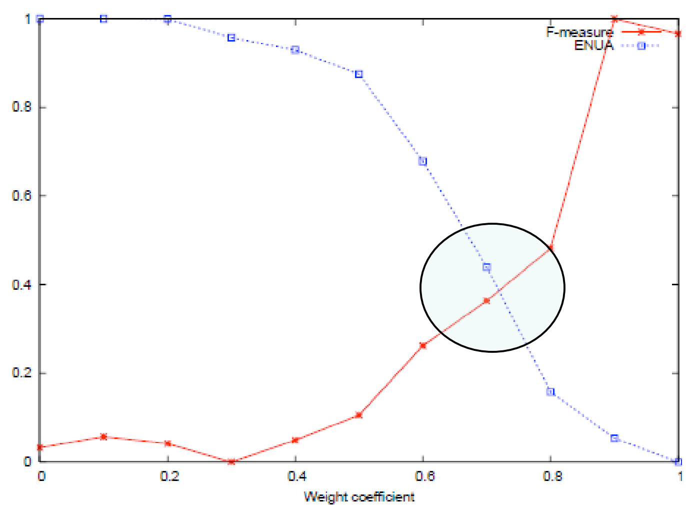
ENUA

- AC: Avg. annotation confidence
- RBT: Rel. docs below threshold
- RDL: Rel. doc length
- AD: Annotation density

Combined measure does better than Random in both cases; marginally for F-Measure & significantly for ENUA ($p < 0.05$)

Conclusion

- There exists a combination of Active & Interactive recommendation strategies which does better for both the measures
- Promising results supporting this claim!



Future Work

- Improvements in Recommendation Strategy
  - Right balance between both measures
  - Automatic estimation of optimal weight combinations
  - Two-stage recommendation
  - Presentation order
- Design the annotation simulator to be more human like
- Calculation of actual number of user actions (ANUA)
  - Analysis of correlation between ENUA & ANUA
- Other recommendation strategies
  - Committee based approaches [McCallum et. al. 1998]
  - Complex annotation types

References

[Culottat et. al., 2005] Culotta, A., Viola, P., & McCallum, A. (2005). Reducing labeling effort for structured prediction tasks. Proc. of AAAI-2005

[Kristjannson et al., 2004] Kristjannson T., Culotta A., Viola P., McCallum A., Interactive information extraction with constrained conditional random fields. In proceedings of AAAI, 2004

[McCallum et al., 1998] McCallum, A. K., Nigam, K., 1998. Employing EM in pool-based active learning for text classification. Proceedings of ICML-98, 15th International Conference on Machine Learning (pp.350358). Madison, US: Morgan Kaufmann Publishers, San Francisco, US.

[Nyberg et al., 2007] Nyberg E., Arora S., Pathak M., Lambert B., Alavedra J., Interactive Annotation Learning :Active Learning for Real-World Text Analysis.(Unpublished Manuscript)

[Sang et al., 2003] Sang E, De Meulder F., 2003. Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition. In proceedings of CoNLL, 2003.(http://www.cnts.ua.ac.be/conll2003/ner)

[Sarawagi et. al., 2002] Sarawagi S. and Bhamidipaty A. 2002. Interactive deduplication using active learning. In Proceedings of the 8th ACM SIGKDD Conference on Knowldge Discovery and Data Minning, Edmonton, Alberta, Canada.

[Seokhwan et. al., 2006] Seokhwan K.., Song Y., Kim K., Cha J., Lee G. G., MMR-based active machine learning for Bio named entity recognition, Proceedings of the HLT-NAACL06, June 2006, New York.

[Shen et al., 2004] Shen D., Zhang J., Su J., Zhou G., Tan C., Multi-Criteria-based Active Learning for Named Entity Recognition. In proceedings of ACL, 2004.

[Thompson et al., 1999] Thompson, C. A., Califf, M. E., Mooney, R. J., 1999. Active learning for natural language parsing and information extraction. In proc. 16th International Conf. on Machine Learning. MorganKaufmann, San Francisco, CA, pp. 406-414.

Questions ?