Bringing Active Learning to Life

Ines Rehbein, Josef Ruppenhofer & Alexis Palmer

COLING 2010
Outline

1. Active Learning
   - Active Learning – Overview
   - Related Work

2. Annotation Experiment
   - Experimental Set-Up
   - Results

3. Conclusion and Future Work

Bringing Active Learning to Life
Outline

1. Active Learning
   - Active Learning – Overview
   - Related Work

2. Annotation Experiment
   - Experimental Set-Up
   - Results

3. Conclusion and Future Work
What this talk is about

- Active Learning (AL): method for limiting the amount of time and cost for human annotation
- AL has successfully been applied to a variety of NLP tasks in the context of supervised machine learning
- Underlying idea:
  - select new training instances with respect to the information content they provide for the machine learning classifier
  - → small number of carefully selected instances can yield as good results as a larger set of randomly chosen examples
- Caveat: most studies have simulated AL using \textit{perfect gold standard data}

Does AL work in a real-world scenario with noisy data?
Active learning loop:
1. select new instances where classifier confidence is below threshold (uncertainty sampling)
2. present them to oracle for manual annotation
3. add them to seed and re-train
continue with (1)
Applying Active Learning to various NLP tasks like

- **POS tagging** (Dagan and Engelson, 1995; Ringger et al., 2007)
- **Parsing** (Hwa, 2004; Baldridge and Osborne, 2004)
- **Named Entity Resolution** (Shen et al., 2004; Tomanek et al., 2007)
- **Word Sense Disambiguation** (Chen et al., 2006; Zhu and Hovy, 2007)
- **Text categorisation** (Lewis and Gale, 1994)
- ...
Nearly all studies have run simulations using *perfect* data.

Fine for tasks with high annotation accuracy and high Inter-Annotator Agreement (IAA).

But: What can we learn from simulations on perfect data when dealing with:
- error-prone data
- inconsistent annotations (low IAA)

Example:
Word Sense Disambiguation of highly ambiguous, fine-grained word meanings.
Put AL to the test - apply it to a real-world scenario

- **SALSA** – The **SA**arbrücken **L**exical **S**emantics **A**cquisition Project
  - Semantic annotation of German newspaper text with semantic frames and roles (Fillmore 1976, 1977, ...)

- The task of frame assignment can be treated as a Word Sense Disambiguation task (Erk, 2005)
Put AL to the test - apply it to a real-world scenario

- **SALSA** – The **SAarbrücken Lexical Semantics Acquisition** Project
  - Semantic annotation of German newspaper text with semantic frames and roles (Fillmore 1976, 1977, ...)

- The task of frame assignment can be treated as a Word Sense Disambiguation task (Erk, 2005)

Many others RUSHED back Wednesday morning.
Put AL to the test - apply it to a real-world scenario

- **SALSA** – The **SAarbrücken Lexical Semantics Acquisition Project**
  - Semantic annotation of German newspaper text with semantic frames and roles (Fillmore 1976, 1977, ...)

- The task of frame assignment can be treated as a Word Sense Disambiguation task (Erk, 2005)

  SELF_MOTION?    FLUIDIC_MOTION?

  Many others    RUSHED  back Wednesday morning.
Put AL to the test - apply it to a real-world scenario

**SALSA** – The **SAarbrücken Lexical Semantics Acquisition** Project

- Semantic annotation of German newspaper text with semantic frames and roles (Fillmore 1976, 1977, ...)

The task of frame assignment can be treated as a Word Sense Disambiguation task (Erk, 2005)

SELF_MOTION?  FLUIDIC_MOTION?

Many others **RUSHED** SELF_MOTION back Wednesday morning.
Outline

1. Active Learning
   - Active Learning – Overview
   - Related Work

2. Annotation Experiment
   - Experimental Set-Up
   - Results

3. Conclusion and Future Work
Annotation Set-Up

- **Annotation** of 8 German causation nouns (Part of on-going SALSA annotation)

- **6 Annotators** (2 authors + 4 computational linguistics undergraduates with at least 1 year experience in frame-semantic annotation)

- **Annotation process**: choosing the correct frame for a target word in context from a list
Targets exhibit a range of difficulty in terms of frame number and ambiguity

| target           | # frames | Fleiss’ \( \kappa \) |
|------------------|----------|-----------------------|
| Anlass occasion  | 2        | 0.67                  |
| Motiv motive     | 2        | 0.79                  |
| Konsequenz consequence | 4    | 0.55                  |
| Quelle source of experience | 6    | 0.77                  |
| Ergebnis effect  | 8        | 0.63                  |
| Resultat result  | 8        | 0.59                  |
| Ausgang outcome  | 9        | 0.82                  |
| Grund reason     | 16       | 0.43                  |
Experimental Setup

- Experimental data:
  - 300 sentences for each target word, split up in packages of 50 sentences each (2,400 instances per annotator, 14,400 total)

- Setup:
  - GUI for annotation, target words highlighted, annotators could use keyboard shortcuts
  - For each instance: record annotation time
  - Control for training effects, switch annotation setting (AL/random) for each package
  - Annotators were not aware of which setting they were in
Dem Wunsch der Preisgründer, die Möglichkeiten der Bild-Text-Kombination speziell für das neue Medium auszuloten, zeitigte erstaunlich hochqualitative Resultate. Unknown1

Da die Zahl der Abonnenten die Marke von 1,4 Millien überschritten habe, könne ein positives Resultat für 1997 erwartet werden. Unknown6

Laut Umfrage des Meinungsforschungsinstituts GlobeScan würde in Deutschland der Herausforderer Kerry gegen Bush mit 74 zu 10 Prozent gewinnen, während das Resultat in Frankreich 64 zu 5 Prozent lauten würde. Unknown7

In beiden Fällen lieferte keine der 34 Gemeinden ein dem Gesamtergebnis widersprechendes Resultat. Unknown7

Dann aber gab Tomba bei der Begrüssung der spanischen Presse noch dem Wunsch Ausdruck, dass ein Vertreter aus dem Veranstalterland im Riesenslalom ein gutes Resultat erzielen möge. Unknown7

Denn die Tatsache, dass das Resultat die Stärkeverhältnisse der Kontrahenten im Zentralstadion recht akkurat spiegelte, vermag nicht darüber hinwegzutauchen, dass die Equipe von Verbandscoach Enzo Trossero die Partie hätte gewinnen können, um nicht zu sagen müssen. Unknown7

Das Resultat vom Spiel in Moskau, 3:1 für Lok, steht, Tirol müsste also 2:0 oder höher gewinnen, um in die Champions League aufzusteigen und nicht wie bereits ausgelost am UEFA-Cup teilzunehmen.
Experimental Setup

- **Experimental data:**
  - 300 sentences for each target word, split up in packages of 50 sentences each (2,400 instances per annotator, 14,400 total)

- **Setup:**
  - GUI for annotation, target words highlighted, annotators could use keyboard shortcuts
  - For each instance: record annotation time
  - Control for training effects, switch annotation setting (AL/random) for each package
  - Annotators were not aware of which setting they were in
Experimental Setup II

- **Random Setting**
  - All annotators had to annotate the same set of randomly selected sentences (150 instances for each target)
  - → compare annotation time and IAA between the annotators

- **Active Learning**
  - 3 pools of 2000 sentences from which the classifier could pick new instances during AL
  - On any AL trial: each annotator uses the same pool of 2000 sentences as all other annotators
  - → in an AL simulation with fixed parameters this would result in the same subset of sentences selected by the classifier
  - → for human annotators, due to different annotation decisions the resulting set of sentences is expected to differ
Experimental Setup III

- **Classifier**
  - MaxEnt (http://maxent.sourceforge.net)

- **Sampling method**
  - uncertainty sampling (Lewis and Gale, 1994)
    - choose instance where classifier confidence is low

- **In each iteration:**
  - select 1 new instance from the pool and present it to the oracle
  - retrain classifier
Does AL speed up the annotation process when working with noisy data?

- How long did it take to assign labels on randomly selected instances compared to the ones selected by Active Learning?
- How many instances do we need to achieve the best (or sufficient) classifier performance for each setting?
Does AL speed up the annotation process when working with noisy data? II

- Overall, annotating the same number of instances on average takes longer in the AL setting
- Effect not due to sentence length (Spearman Rank correlation)
So AL doesn’t work here?

- Instances selected by Active Learning took longer to annotate
- But what about classifier performance?

**Ausgang, Motiv:**
AL gives a substantial boost in classifier performance of 5% and 7% accuracy

**Konsequenz, Quelle:**
somewhat smaller gains of 2% and 1%

**Anlass, Ergebnis, Resultat, Grund:**
same accuracy or even worse
Learning Curves for Results

**Random Sampling**

**Active Learning**

- **No. of iterations**
- **Accuracy**
- **Annotator**: A1, A2, A3, A4, A5, A6

Rehbein, Ruppenhofer & Palmer

Bringing Active Learning to Life
Learning Curves for Konsequenz

Random Sampling

Active Learning

Accuracy

No. of iterations

No. of iterations

0 50 100 150

0.3 0.4 0.5 0.6 0.7

0 50 100 150

0.3 0.4 0.5 0.6 0.7

Rehbein, Ruppenhofer & Palmer

Bringing Active Learning to Life
Learning Curves for Motiv

Random Sampling

Active Learning

No. of iterations

Accuracy

No. of iterations

Accuracy
Why does AL work for some targets but not for others?

Explanation 1: Inter-Annotator Agreement?

| target    | $\kappa$ | Acc. |
|-----------|----------|------|
| + Ausgang | 0.82     | 0.94 |
| + Motiv   | 0.79     | 0.92 |
| + Quelle  | 0.77     | 0.88 |
| - Anlass  | 0.67     | 0.88 |
| - Ergebnis| 0.63     | 0.76 |
| - Resultat| 0.59     | 0.75 |
| + Konsequenz | 0.55   | 0.80 |
| - Grund   | 0.43     | 0.72 |

⇒ Ok for Ausgang, Motiv, Quelle – why Konsequenz?
Why does AL work for some targets but not for others?

Explanation 2: Accuracy of annotations?

| target  | $\kappa$ | Acc. |
|---------|----------|------|
| + Ausgang | 0.82 | 0.94 |
| + Motiv   | 0.79 | 0.92 |
| + Quelle  | 0.77 | 0.88 |
| - Anlass  | 0.67 | 0.88 |
| - Ergebnis | 0.63 | 0.76 |
| - Resultat | 0.59 | 0.75 |
| + Konsequenz | 0.55 | 0.80 |
| - Grund   | 0.43 | 0.72 |

$\Rightarrow$ Ok for Ausgang, Motiv, Quelle, Konsequenz – why not Anlass?
Why does AL work for some targets but not for others?

Explanation 3: Impact of most frequent frame / no. of frames?

| target     | $\kappa$ | Acc. | most freq. | # frames |
|------------|----------|------|------------|----------|
| ++ Ausgang | 0.82     | 0.94 | 0.67       | 9        |
| ++ Motiv   | 0.79     | 0.92 | 0.53       | 2        |
| ++ Quelle  | 0.77     | 0.88 | 0.56       | 6        |
| - Anlass   | 0.67     | 0.88 | 0.63       | 2        |
| - Ergebnis | 0.63     | 0.76 | 0.36       | 8        |
| - Resultat | 0.59     | 0.75 | 0.36       | 8        |
| ++ Konsequenz | 0.55   | 0.80 | 0.61       | 4        |
| - Grund    | 0.43     | 0.72 | 0.58       | 16       |

$\Rightarrow$ No correlation
Correlation between individual annotators and AL

- AL seems to work well for one particular annotator (A6)
- Possible explanations for relation between individual annotators and AL:
  
  (a) annotation accuracy of individual annotators
  (b) quality of training instances
  (c) distribution of frames in the individual data sets for each annotator
Do differences between annotation accuracy for individual annotators influence the AL process?

| Konsequenz | A1 | A2 | A3 | A4 | A5 | A6 |
|------------|----|----|----|----|----|----|
| human      | 0.80 | 0.72 | 0.89 | 0.73 | 0.89 | 0.76 |
| classifier | 0.60 | 0.63 | 0.67 | 0.60 | 0.63 | 0.64 |

no strong correlation between the accuracy of the human annotations and the performance of the classifier trained on these annotations
Active Learning and the Quality of Individual Training Instances

- Are the training sets selected by Annotator 6 (A6) more helpful for the classifier?
- Take the instances from A6 (good training set) and give them to the other annotators for re-annotation.
- Re-train the classifiers and compare performance.
Active Learning

Experimental Set-Up

Results

Active Learning

No. of iterations

Accuracy

Annotator
A1
A2
A3
A4
A5
A6

Re–annotated samples

No. of iterations

Accuracy

Rehbein, Ruppenhofer & Palmer
Bringing Active Learning to Life
Does AL alleviate the class imbalance problem by extracting more balanced data sets? (Ertekin, 2007)

| Ergebnis  | Frame         | R (%) | AL (%) |
|-----------|---------------|-------|--------|
| Causation | 4.8           | 3.7   |
| Outcome   | 17.8          | 10.5  |
| Finding_out | 26.2       | 8.2   |
| Efficacy  | 0.8           | 0.1   |
| Decision  | 5.1           | 3.2   |
| Mathematics | 1.6         | 0.4   |
| Operating_result | 24.5 | 66.7  |
| Competitive_score | 19.2 | 7.2   |

Classifier had problems learning one particular sense → skewed distribution in the AL data

AL did not work for this particular target!
Active Learning and the Quality of Individual Training Instances 2

- Active Learning success not directly dependent on accuracy of annotations
- Particular instances are more informative for the classifier than others
- Great impact of distribution of frames on classifier performance
Conclusions

- We showed that annotation noise can mislead the classifier and result in skewed data sets.
- Under certain conditions, classifier performance for AL can improve over a random sampling baseline even on noisy data.
- Critical features influencing the outcome of AL for this task:
  - the amount of noise in the data
  - the distribution of frames in training- and test sets
- Addressing the class imbalance problem is crucial for applying AL to a real annotation task.
Future Work

- Explore interaction between different factors on Active Learning
  - Inter-Annotator Agreement
  - Accuracy against a gold standard
  - Distribution of frames
  - Number of word senses
  - Majority baseline

- Predict whether AL will work for a specific task or not
Thank You!

Questions?
|        | Random |        |        | Uncertainty |        |        |
|--------|--------|--------|--------|-------------|--------|--------|
|        | 50     | 100    | 150    | 50          | 100    | 150    |
| Anlass | 0.85   | **0.86** | 0.85   | 0.84        | 0.85   | **0.84** |
| Motiv  | 0.57   | 0.62   | 0.63   | 0.64        | 0.67   | **0.70** |
| Konseq. | 0.55  | 0.59   | 0.60   | 0.61        | **0.62** | **0.62** |
| Quelle | 0.56   | 0.53   | 0.54   | 0.52        | 0.52   | **0.57** |
| Ergebnis | 0.39 | **0.42** | 0.41   | 0.39        | 0.37   | 0.38   |
| Resultat | 0.31 | 0.35   | **0.37** | 0.32        | 0.34   | 0.34   |
| Ausgang | 0.67   | 0.69   | 0.69   | 0.68        | 0.72   | **0.74** |
| Grund  | **0.48** | 0.47   | 0.47   | 0.47        | 0.44   | **0.48** |

**Table:** Avg. classifier performance (acc.) over all annotators for the 8 target lemmas when training on 50, 100 and 150 annotated instances for random samples and uncertainty samples
### Table: Annotation time (sec/instance) per target/annotator/setting and average sentence length (sl)

|      | Anlass   | Motiv     | Konsequenz | Quelle  |
|------|----------|-----------|------------|---------|
|      | R        | U         | R          | U       | R        | U       |
| A₁   | 8.6      | 9.6       | 5.9        | 6.6     | 10.7     | 10.5    | 6.0      | 4.8      |
| A₂   | 4.4      | 5.7       | 4.8        | 5.9     | 8.2      | 9.2     | 4.9      | 4.9      |
| A₃   | 9.9      | 9.2       | 6.8        | 6.7     | 6.8      | 8.3     | 7.4      | 6.1      |
| A₄   | 5.8      | 4.9       | 3.6        | 3.6     | 9.9      | 11.3    | 4.8      | 3.5      |
| A₅   | 3.0      | 3.5       | 3.0        | 2.6     | 4.8      | 4.9     | 3.8      | 3.0      |
| A₆   | 5.4      | 6.3       | 5.3        | 4.7     | 6.7      | 8.6     | 5.4      | 4.6      |
| φ    | 6.2      | 6.5       | 4.9        | 5.0     | 7.8      | 8.8     | 5.4      | 4.5      |
| sl   | 25.8     | 27.8      | 27.8       | 26.0    | 24.2     | 25.8    | 24.9     | 26.5     |

|      | Ergebnis | Resultat  | Ausgang   | Grund   |
|------|----------|-----------|-----------|---------|
|      | R        | U         | R          | U       | R        | U       | R        | U       |
| A₁   | 10.5     | 7.4       | 10.1       | 9.6     | 6.4      | 10.0    | 10.2     | 11.1     |
| A₂   | 6.4      | 4.4       | 11.7       | 8.5     | 5.1      | 7.7     | 9.0      | 9.3      |
| A₃   | 9.4      | 7.6       | 9.0        | 12.3    | 7.5      | 8.5     | 11.7     | 10.2     |
| A₄   | 7.9      | 7.1       | 9.7        | 11.1    | 3.6      | 4.1     | 9.9      | 9.4      |
| A₅   | 6.8      | 4.8       | 6.7        | 6.1     | 3.1      | 3.5     | 6.3      | 6.0      |
| A₆   | 7.8      | 6.1       | 8.7        | 9.0     | 6.9      | 6.6     | 9.3      | 8.5      |
| φ    | 8.1      | 6.2       | 9.3        | 9.4     | 5.4      | 6.7     | 9.4      | 9.1      |
| sl   | 25.7     | 25.2      | 29.0       | 35.9    | 25.5     | 27.9    | 26.8     | 29.7     |

*Table: Annotation time (sec/instance) per target/annotator/setting and average sentence length (sl)*