Introduction to Machine Learning
NPFL 054
http://ufal.mff.cuni.cz/course/npfl054

Barbora Hladká       Martin Holub

{Hladka | Holub}@ufal.mff.cuni.cz

Charles University,
Faculty of Mathematics and Physics,
Institute of Formal and Applied Linguistics
Purpose of the demo task
= to show several things related to gold standard data for a supervised machine learning task, especially

- Manual annotation and basic data analysis
- Gold Standard data distribution
- Inter-annotator agreement
- Confusion matrices
- Error analysis
**Verb Patterns Classification** is a kind of *lexical disambiguation* of verbs. The task is similar to the traditional *word sense disambiguation* (WSD). The two tasks differ in how the semantic categories are defined (word senses vs. patterns of typical verb usage).

Let’s focus on two English verbs, namely *cry* and *enlarge.*
CRY -- dictionary definitions

cry ❆  *****

1  cry; cries; crying; cried
When you cry, tears come from your eyes, usually because you are unhappy or hurt.
   I hung up the phone and started to cry.
   Please don't cry.
   He cried with anger and frustration.
   ...a crying baby.
   VB

2  cry; cries; crying; cried
If you cry something, you shout it or say it loudly.
   `Nancy Drew,' she cried, `you're under arrest!'.
   I cried: `It's wonderful news!'
   VB

5  cry; cries
You can refer to a public protest about something or appeal for something as a cry of some kind. (JOURNALISM)
   There have been cries of outrage about this expenditure.
   Many other countries have turned a deaf ear to their cries for help.
   N-COUNT: usu N of/for n
enlarge

1 When you **enlarge** something or when it **enlarges**, it becomes bigger.
   ...the plan to **enlarge** Ewood Park into a 30,000 all-seater stadium...
   The glands in the neck may **enlarge**.

V-ERG

2 To **enlarge** a photograph means to develop a bigger print of it.
   ...newly-weds wishing to **enlarge** snaps of their big day.

VB

3 If you **enlarge** on something that has been mentioned, you give more details about it. *(FORMAL)*
   He didn't **enlarge** on the form that the interim government and assembly would take.
   I wish to **enlarge** upon a statement made by Gary Docking.

VB

= expand
CRY -- Pattern definitions

**Pattern 1**  
[Human] cry [no object]

Explanation  
[[Human]] weeps  
usually because [[Human]] is unhappy or in pain

Example  
*His advice to stressful women was: `If you cry, don't cry alone.*

**Pattern 4**  
[Human] cry [THAT-CL|WH-CL|QUOTE] ({out})

Explanation  
[[Human]] shouts ([QUOTE]) loudly  
typically, in order to attract attention

Example  
*You can hear them screaming and banging their heads, crying that they want to go home.*

**Pattern 7**  
[Entity | State] cry [{out}] [{for} Action] [no object]

Explanation  
[[Entity | State]] requires [[Action]] to be taken urgently

Example  
*Identifying areas which cry out for improvement or even simply areas of muddle and misunderstanding, is by no means negative -- rather a spur to action.*
**ENLARGE -- Pattern definitions**

**Pattern 1**  
[[Human]^[[Eventuality]]] enlarge [Entity]

Explanation  
[[Human | Eventuality]] causes [[Entity]] to grow or become larger

Example  
*These were not large powers, but later changes were to **enlarge** them.*

**Pattern 2**  
[Entity] enlarge [no object]

Explanation  
[[Entity]] grows or becomes larger

Example  
*As infants grow, their bodies not only **enlarge** but change both in shape and colour.*

**Pattern 3**  
[[Human]^[[Document]]] enlarge [[on | upon} Anything = Topic] [no object]

Explanation  
[[Human]] speaks or writes at length on [[Anything = Topic]] or [[Document]] contains long-winded comments on [[Topic]]

Example  
*Let me **enlarge** on this a little.*

**Pattern 4**  
enlarged

Explanation  
now larger than before, without any deliberate causer or causer irrelevant

Example  
*The fluid filled spaces or ventricles appear to be **enlarged**, and the blood flow to the front of the brain is reduced.*
Verb Patterns Classification – annotation description

You will classify *cry* and *enlarge* manually.

- You will be given 10+10 sentences with the given verbs.
- For each sentence you will assign a pattern that fits best the given sentence.
  - There are 3 predefined patterns for the verb *cry*.
  - There are 4 predefined patterns for the verb *enlarge*.
  - If you think that no pattern matches the sentence, choose "u".
  - If you think that the given word is not a verb, choose "x".
- Use the forms posted at https://ufal.mff.cuni.cz/courses/npfl054/demo
Gold standard data sets are posted on the course web page (DEMO).

CRY – 250 instances in the GS set

<table>
<thead>
<tr>
<th>class</th>
<th>1</th>
<th>4</th>
<th>7</th>
<th>u</th>
<th>x</th>
</tr>
</thead>
<tbody>
<tr>
<td>frequency</td>
<td>131</td>
<td>59</td>
<td>13</td>
<td>33</td>
<td>14</td>
</tr>
</tbody>
</table>

ENLARGE – 300 instances in the GS set

<table>
<thead>
<tr>
<th>class</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>u</th>
</tr>
</thead>
<tbody>
<tr>
<td>frequency</td>
<td>230</td>
<td>21</td>
<td>20</td>
<td>26</td>
<td>3</td>
</tr>
</tbody>
</table>
**Automatic classifier** is a function that assigns certain output class to each input instance.

**Output class** is a discrete (possibly categorical) value.

In the demo task: Pattern tags are categorical output values, sentences containing the verbs in question are input instances.

**Classifier accuracy** is often estimated using a test data sample as a percentage of correctly classified instances in the sample. This estimate is called *sample accuracy*.

**Automatic predictions** made by automatic classifier (our best model F1) are posted on the course web page (DEMO).

– NOTE that it is the same GS set, and it was also used as training data (!).
– Thus, you can compute only the training error, not the test error.
Manual annotation

Annotated data – a subset of the GS
– the same data set annotated by each group

2014 – 2 groups
- A (5 Czech)
- B (2 Czech, 3 foreign)

2015 – 4 groups
- A (6 Czech)
- B (6 Czech)
- C (6 Czech)
- D (6 Czech)

Now we can analyse/compare
- which group is closer to the Gold Standard
- inter-annotator agreement between groups
- error types
  - made by people
  - made by automatic classifier
A, B and GS distributions - CRY (2014)
A vs GS - confusion matrix - CRY (2014)

Cry A histogram

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>4</th>
<th>7</th>
<th>u</th>
<th>x</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>27</td>
<td>2</td>
<td>0</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>6</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>A</td>
<td>7</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>u</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>x</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>

Number of agreements: 35 (70%)
Number of disagreements: 15 (30%)

<table>
<thead>
<tr>
<th></th>
<th>GS</th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>NPFL054, 2017 Hladká &amp; Holub Demo 1, page 15/29</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>A</strong></td>
<td>u</td>
<td>x 1</td>
<td>4</td>
<td>7</td>
<td>3</td>
<td>Agreement: 41 (68%) Disagreement: 19 (32%)</td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>0</td>
<td>29</td>
<td>3</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>1</td>
<td>0</td>
<td>5</td>
<td>8</td>
<td></td>
</tr>
<tr>
<td></td>
<td>7</td>
<td>3</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td><strong>B</strong></td>
<td>u</td>
<td>x 1</td>
<td>4</td>
<td>7</td>
<td>3</td>
<td>Agreement: 40 (67%) Disagreement: 20 (33%)</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>1</td>
<td>3</td>
<td>2</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>30</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>2</td>
<td>0</td>
<td>3</td>
<td>6</td>
<td></td>
</tr>
<tr>
<td></td>
<td>7</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td><strong>C</strong></td>
<td>u</td>
<td>x 1</td>
<td>4</td>
<td>7</td>
<td>3</td>
<td>Agreement: 40 (67%) Disagreement: 20 (33%)</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>0</td>
<td>5</td>
<td>2</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>28</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>1</td>
<td>0</td>
<td>3</td>
<td>6</td>
<td></td>
</tr>
<tr>
<td></td>
<td>7</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td><strong>D</strong></td>
<td>u</td>
<td>x 1</td>
<td>4</td>
<td>7</td>
<td>3</td>
<td>Agreement: 45 (75%) Disagreement: 15 (25%)</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>0</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>1</td>
<td>3</td>
<td>1</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>0</td>
<td>1</td>
<td>4</td>
<td>10</td>
<td></td>
</tr>
<tr>
<td></td>
<td>7</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td></td>
</tr>
</tbody>
</table>
Agreement: 38 (63 %)
Disagreement: 22 (37 %)

Agreement: 28 (47 %)
Disagreement: 32 (53 %)

Agreement: 28 (47 %)
Disagreement: 32 (53 %)

Agreement: 36 (60 %)
Disagreement: 24 (40 %)
Inter-annotator agreement (IAA) (2014)

**CRY** – confusion matrix (50 instances, 33 agreements = 66%)

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>4</th>
<th>7</th>
<th>u</th>
<th>x</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>24</td>
<td>3</td>
<td>1</td>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>3</td>
<td>3</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

**ENLARGE** – confusion matrix (50 instances, 31 agreements = 62%)

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>u</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>18</td>
<td>2</td>
<td>0</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>4</td>
<td>7</td>
<td>1</td>
<td>4</td>
<td>0</td>
</tr>
</tbody>
</table>

Hladká & Holub Demo 1, page 17/29
What agreement would be reached by chance?

**Example 1**

Assume two annotators \((A_1, A_2)\), two classes \((t_1, t_2)\), and the following distribution:

<table>
<thead>
<tr>
<th></th>
<th>(t_1)</th>
<th>(t_2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(A_1)</td>
<td>50 %</td>
<td>50 %</td>
</tr>
<tr>
<td>(A_2)</td>
<td>50 %</td>
<td>50 %</td>
</tr>
</tbody>
</table>

Then

- the best possible agreement is 100 %
- the worst possible agreement is 0 %
- the “agreement-by-chance” would be 50 %
Example 2

Assume two annotators \((A_1, A_2)\), two classes \((t_1, t_2)\), and the following distribution:

<table>
<thead>
<tr>
<th></th>
<th>(t_1)</th>
<th>(t_2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(A_1)</td>
<td>90%</td>
<td>10%</td>
</tr>
<tr>
<td>(A_2)</td>
<td>90%</td>
<td>10%</td>
</tr>
</tbody>
</table>

Then

- the best possible agreement is 100%
- the worst possible agreement is 80%
- the “agreement-by-chance” would be 82%
What agreement would be reached by chance?

Example 3

Assume two annotators ($A_1$, $A_2$), two classes ($t_1$, $t_2$), and the following distribution:

<table>
<thead>
<tr>
<th></th>
<th>$t_1$</th>
<th>$t_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$A_1$</td>
<td>90 %</td>
<td>10 %</td>
</tr>
<tr>
<td>$A_2$</td>
<td>80 %</td>
<td>20 %</td>
</tr>
</tbody>
</table>

Then

- the best possible agreement is 90%
- the worst possible agreement is 70%
- the “agreement-by-chance” would be 74%
The situation from Example 3 can be simulated in R

```r
# N will be the sample size
> N = 10^6

# two annotators will annotate randomly
> A1 = sample(c(rep(1, 0.9*N), rep(0, 0.1*N)))
> A2 = sample(c(rep(1, 0.8*N), rep(0, 0.2*N)))

# percentage of their observed agreement
> mean(A1 == A2)
[1] 0.740112

# exact calculation -- just for comparison
> 0.9*0.8 + 0.1*0.2
[1] 0.74
```
Cohen’s kappa was introduced by Jacob Cohen in 1960.

\[ \kappa = \frac{\Pr(a) - \Pr(e)}{1 - \Pr(e)} \]

- \( \Pr(a) \) is the relative observed agreement among annotators
  = percentage of agreements in the sample
- \( \Pr(e) \) is the hypothetical probability of chance agreement
  = probability of their agreement if they annotated randomly
- \( \kappa > 0 \) if the observed agreement is better than what would be expected by chance

Limitations

- Cohen’s kappa measures agreement between two annotators only
- for more annotators you should use the more general Fleiss’ kappa
  – see http://en.wikipedia.org/wiki/Fleiss’_kappa
Inter-annotator agreement (2014)

CRY
Number of agreements: 33 (66 %)
Number of disagreements: 17 (34 %)
Cohen’s kappa: 0.437
Fleiss’s kappa: 0.434

ENLARGE
Number of agreements: 31 (62 %)
Number of disagreements: 19 (38 %)
Cohen’s kappa: 0.438
Fleiss’s kappa: 0.433
Inter-annotator agreement (2015)

### CRY – Cohen’s kappa

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>–</td>
<td>0.36</td>
<td>0.28</td>
<td>0.41</td>
</tr>
<tr>
<td>B</td>
<td>–</td>
<td>–</td>
<td>0.37</td>
<td>0.41</td>
</tr>
<tr>
<td>C</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>0.33</td>
</tr>
<tr>
<td>D</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
</tbody>
</table>

### ENLARGE – Cohen’s kappa

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>–</td>
<td>0.31</td>
<td>0.41</td>
<td>0.30</td>
</tr>
<tr>
<td>B</td>
<td>–</td>
<td>–</td>
<td>0.22</td>
<td>0.32</td>
</tr>
<tr>
<td>C</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>0.37</td>
</tr>
<tr>
<td>D</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
</tbody>
</table>

CRY – **Fleiss’s kappa** 0.35

ENLARGE – **Fleiss’s kappa** 0.32
### Automatic classifier – training error analysis

**ENLARGE (2014)**

<table>
<thead>
<tr>
<th>GS</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>u</th>
<th>GS</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>u</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>224</td>
<td>1</td>
<td>1</td>
<td>12</td>
<td>2</td>
<td>1</td>
<td>0.97</td>
<td>0.05</td>
<td>0.05</td>
<td>0.46</td>
<td>0.67</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>17</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>0.01</td>
<td>0.81</td>
<td>0.15</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>2</td>
<td>15</td>
<td>0</td>
<td>0</td>
<td>3</td>
<td>0.00</td>
<td>0.10</td>
<td>0.75</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>4</td>
<td>3</td>
<td>1</td>
<td>0</td>
<td>14</td>
<td>1</td>
<td>4</td>
<td>0.01</td>
<td>0.05</td>
<td>0.00</td>
<td>0.54</td>
<td>0.33</td>
</tr>
<tr>
<td>u</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>u</td>
<td>0.00</td>
<td>0.00</td>
<td>0.05</td>
<td>0.00</td>
<td>0.00</td>
</tr>
</tbody>
</table>

**Number of agreements:** 270 (90%)

**Number of disagreements:** 30 (10%)
<table>
<thead>
<tr>
<th></th>
<th>GS</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th>GS</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>u</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>u</td>
</tr>
<tr>
<td>A+B</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>A+B</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>46</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0.64</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>2</td>
<td>11</td>
<td>14</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>2</td>
<td>0.15</td>
<td>1.00</td>
<td>0.00</td>
<td>0.08</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>3</td>
<td>0.04</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>4</td>
<td>12</td>
<td>0</td>
<td>0</td>
<td>10</td>
<td>0</td>
<td>4</td>
<td>0.17</td>
<td>0.00</td>
<td>0.00</td>
<td>0.83</td>
</tr>
<tr>
<td>u</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>u</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.08</td>
</tr>
</tbody>
</table>

**Number of agreements:** 72 (72%)
**Number of disagreements:** 28 (28%)
Summary of Lab #1
Examination Requirements

You should be able to practically compute and understand/use

- categorical data distribution
- confusion matrices
- classifier accuracy
- inter-annotator agreement
  - simple percentage
  - Cohen’s kappa
- probability (both conditional and unconditional) of errors of different types
Practical exercises in R

- Download two files with annotated data cry-A.csv and cry-C.csv.
  - [https://ufal.mff.cuni.cz/courses/npfl054/demo](https://ufal.mff.cuni.cz/courses/npfl054/demo)

- Run R and read the data using `read.csv()`.
  - Hint: see the posted Tutorial, Part I.
  - ... and create objects `cry.A` and `cry.C`.

- Make the confusion matrix between groups A and C.
  - Hint: use `table(cry.A$class, cry.C$class)`

- Compute simple agreement (in percentage) between A and C.
  - Hint: use `diag()` and `sum()`

- Compute the Cohen’s kappa value between groups A and C.
  - For hints see Part III of the Tutorial.
Summary of Lab #1

Homework

• Go through all details in the Tutorial (Parts I, II, and III)

• Get familiar with the `data.table` package
  – just to understand Part II

• Do all exercises in Part III