# Exam INF5830, 2015, Solutions

## 1 Evaluation and significance (20%)

(a) You are testing two classifiers for supervised relation extraction. You are testing them on 100 labeled test items. Classifier 1 classfies 60 of the items correctly, while classifier 2 classifies 50 of the items correctly. Would you from this observation conclude that classifier 1 is significantly better than classifier 2? State reasons for your answer.

Using the two-sample "t-test" as it applies to proportions, yields a $z$-score of

$$z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}} = \frac{0.6 - 0.5}{\sqrt{\frac{0.6(1-0.6)}{100} + \frac{0.5(1-0.5)}{100}}} =$$

$$\frac{0.1}{\sqrt{\frac{0.24+0.25}{100}}} = \frac{10}{7} < 1.5$$

Hence it is not significant at the 0.05 level.

(b) Suppose that you in addition for each of the 100 test items record the result of each of the two classifiers, and get the following numbers:

- Both classifiers are correct on 45 items.
- Both classfiers are incorrect on 35 items.
- Classifier 1 is correct and classifier 2 is incorrect on 15 items.
- Classifier 2 is correct and classifier 1 is incorrect on 5 items.

Would you from these observations conclude that classifier 1 is significantly better than classifier 2? State reasons for your answers. (In case you find the actual calculations hard, explain how you would proceed to solve the exercise if you had a computer or calculator available.)

**Alternative 1: Sign-test** We only compare the 20 items where the two classifiers disagree. If the two classifiers were equally good, there should be a chance of $p_0 = 0.5$ that classifier 1 is correct and classifier 2 is incorrect. Since this happens 15 out of 20, we have the observation $\hat{p} = \frac{20}{15} = 0.75$. How unlikely is it to get 15 or more out of 20? With a computer we could use the binomial distribution to calculate this. Without a computer, we can use the normal distribution to approximate the binomial distribution since $p_0 n = (1 - p_0)n = 10$.

$$z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} = \frac{0.75 - 0.5}{\sqrt{\frac{0.5(1-0.5)}{20}}} = \frac{0.25}{\frac{0.5}{\sqrt{20}}} = \sqrt{5} > 2.2$$

Since this is larger than 1.645, classifier 1 is significantly better than classifier 2 at the 0.05 level.

**Alternative 2: t-test for matched pairs** Use 1 for correct and 0 for incorrect and let $x$ be a variable which for each test item records the difference between classifier 1 and classifier 2. Then

$$\bar{x} = \frac{45 \times (1-1) + 35 \times (0-0) + 15 \times (1-0) + 5 \times (0-1)}{100} = 0.1$$

$$s^2 = \frac{80 \times (0-0.1)^2 + 15 \times (1-0.1)^2 + 5 \times (-1-0.1)^2}{99} =$$

$$= \frac{80 \times 0.01 + 15 \times 0.81 + 5 \times 1.21}{99} = \frac{0.8 + 12.15 + 6.05}{99} = \frac{19}{99}$$

We are using a one-sided $t$-test to see whether this is significantly larger than 0.

$$t = \frac{\bar{x} - \mu}{\sqrt{\frac{s^2}{n}}} = \frac{0.1 - 0}{\sqrt{\frac{\frac{19}{99}}{100}}} = \sqrt{\frac{99}{19}} > \sqrt{\frac{95}{19}} = \sqrt{5} > 2.2$$

We do not have a table for the $t$-distribution with 99 degrees of freedom, but we see that even if we use the distribution for 90 degrees of freedom, this is definitely significant at the 0.05 level (which requires a $t$-score of 1.662).

## 2 Dependency syntax and parsing (30 %)

(a) Draw the dependency graph for the sentence *He takes his wash to the laundromat*, here provided in the so-called CoNLL-format:

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 1 | He | he | PRP | PRP | _ | 2 | SBJ |
| 2 | takes | take | VBZ | VBZ | _ | 0 | ROOT |
| 3 | his | his | PRP | PRP | _ | 4 | NMOD |
| 4 | wash | wash | NN | NN | _ | 2 | OBJ |
| 5 | to | to | TO | TO | _ | 2 | DIR |
| 6 | the | the | DT | DT | _ | 7 | NMOD |
| 7 | laundromat | laundromat | NN | NN | _ | 5 | PMOD |

Draw graph

(b) Provide two examples of a formal condition on dependency graphs. Draw a version of the graph in a) that violates these conditions.
Two from the following:

- single-headedness: every word has at most one head
  - If $i \rightarrow j$, then not $k \rightarrow j$, for any $k \neq i$.
- acyclicity: no cycles, structure is hierarchical
  - If $i \rightarrow j$ then not $j \rightarrow^* i$.

- connectedness: every word has a head, structure is complete (enforced by adding a dummy root node)
  - For every node $i$ there is a node $j$ such that $i \rightarrow j$ or $j \rightarrow i$.
- projectivity: no crossing branches
  - If $i \rightarrow j$ then $i \rightarrow^* k$, for any $k$ such that $i<k<j$ or $j<k<i$.

+ drawing of graph

(c) Nivre's arg-eager algorithm is a commonly used algorithm for transition-based dependency parsing. Describe the four transitions employed in Nivre's algorithm. Your description should make reference to the stack and queue data structures as well as the conditions that apply to the application of each transition.

**SHIFT:** top of queue is shifted to top of stack, condition: none

**REDUCE:** top element i of stack is removed / stack is popped, condition: i has a head

**LEFT-ARC:** arc is added from top of queue to top of stack, stack is popped, condition: top of stack does not have a head

**RIGHT-ARC:** arc is added from top of stack to top of queue, top of queue is pushed onto stack, condition: top of queue does not have a head

# 3 Information extraction (30%)

(a) What are the typical steps of an information extraction system? Explain what the **goals** are for each step. You do not have to explain **how** the actual steps are carried out.

The overall goal is to extract information from textual material.

**Clean up** First make sure that the data is interpreted correctly as text, taking into consideration e.g., encoding. May also be necessary to remove—or put aside—meta-data as e.g., XML- or HTML-tags.

**Sentence segmentation** Split the text into units corresponding to sentences.

**Word tokenization** Split each sentencet into a sequence of smaller units corresponding to words. May also be units corresponding to punctuation. Some systems prefer to put some multi-word expressions into one unit.

**Part of speech tagging** For each sentence, tag the words with a part-of-speech tag, e.g., whether the word is a noun or a verb. This includes a disambiguation of the word from the context, as many words may occur in several word classes, e.g., "run" may be either a verb or a noun.

**Chunking** gather words into so-called NP-chunks, e.g.,

(The president) of the (United States) gave (a speech) .

These are maximally large NPs that do not contain other NP chunks. Some chunkers may also identify verbs and preposition chunks, e.g., "gave" as a VP chunk in the example.

**Named entity recognition** To each NP chunk, decide whether it can name an entity and if it can, assign a class from a predefined small set of classes, e.g., **person** which may apply to "the president" in the example, and **organization** which might apply to "the United States". In a different context "the United States" could be classified as a **location**.

**Relation extraction** Extract relations that exist between the named entities of the text. Normally a pre-defined set of relations determined by the purpose of the application, e.g., for medical records, this could include **date of birth**, **has sympton**, **has diagnosis** etc.

**Additional steps** The extraction of temporal expressions and events can be additional steps which may be used to extract e.g., not only that the pasient has a sympton, but when the sympton first appeared.

(b) One of the steps in an information extraction system is relation extraction. There are several different methods for relation extraction. One method is to use hand-written patterns, another method is to apply supervised classification. Explain shortly the main principles of the two approaches. What are the bottlenecks of each approach?

**Hand written patterns** A person tries to identify patterns that are commonly used to express the relation, e.g., *x wrote y*, *x is the author of y* and *y is a book by x* for authorship.

The bottleneck is to write these patterns general enough, and to include sufficiently many of them, as there are often many ways to express the same.

**Supervised classification** Sentences are manually annotated with named entities and relations between them. This is used to train a classifier which can then be used to assign relations to other sentences (after they have been tagged, chunked and named-entity recognized).

The bottleneck is that one needs much training material to get this to work satisfactorily and that it is resource demanding to make this training material.

(c) A third method is to use semi-supervised classfiers that are constructed by so-called bootstrapping. Explain how this method works.

- If we know patterns for a relation, that may be used to find pairs that stand it that relationship, e.g., from "x was written by y" we may extract *Hamlet—Shakespeare* and *Harry Potter—J.K.Rowling*
- Coversely, if we have a set of book-author pairs, we can search for patterns that occur in sentences where the book and author co-occur.

We can use this sytematically switching between the two steps.

1. Say, we start with a set of pairs of books and authors.
2. We gather many sentences each containing a pair.
3. We try to recognize patterns that are recurrent in these sentences.
4. For each pattern, we gather more sentences containing the pattern.
5. From these sentences, we extract new candidate pairs of books and authors.
6. We evaluate the quality of the extracted candidate pairs for each pattern and keep the patterns that yield a satisfactory result.
7. We use these newly acquired patterns to extract more book-author candidates, and repeat from step 2.

The patterns to which we refer here, may have the form of a set of features and values.

# 4 Semantic role labeling (20 %)

We want to construct a Semantic Role Labeling (SRL) system and need to consider several issues in order to do so, in particular we will be considering the available linguistic resources for this task and the architecture of our system.

(a) There are two main resources which include semantic role information for English: FrameNet and PropBank. Provide a short description of each of these and point to their main differences.
In PropBank

- roles defined wrt individual verb senses
  - Arg0 and Arg1 correspond to proto-agent and proto-patient (Dowty)
  - Arg2, Arg3 etc numbered roles specific for verb sense
  - non-numbered arguments (ArgMs) represent adjunct roles, e.g. temporal, locative, directional etc.
- is a corpus which labels all sentences in Penn Treebank with semantic roles according to this scheme

FrameNet:

- roles specific to a frame (semantic situation/background knowledge/model in AI)
- frame often covers several predicates
- core vs non-core roles (like ArgM)
- both computational lexicon and annotated corpus (but small and manually selected, not representative)

The main difference between the two is found in the the set of roles: verb sense specific + set of general adjuncts roles (PropBank) and frame-specific (FrameNet) Another difference is in the type of corpora available for English: the whole PTB (PropBank) vs a small corpus of manually selected sentences.

(b) In the following, the earlier sentence has been annotated with semantic roles in the (somewhat abbreviated) CoNLL08 format:

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 1 | He | ... | _ | 2 | SBJ | _ | A0 |
| 2 | takes | ... | _ | 0 | ROOT | take.01 | _ |
| 3 | his | ... | _ | 4 | NMOD | _ | _ |
| 4 | wash | ... | _ | 2 | OBJ | _ | A1 |
| 5 | to | ... | _ | 2 | DIR | _ | AM-DIR |
| 6 | the | ... | _ | 7 | NMOD | _ | _ |
| 7 | laundromat | ... | _ | 5 | PMOD | _ | _ |

(i) Extract the predicate and the semantic arguments along with their roles from the above example. Provide a short description for each of the roles.

* Predicate: take.01
* Arguments:
  - He : A0 : Agent, doer
  - his wash : A1 : Patient, affected
  - to the laundromat : AM-DIR : directional adjunct

(ii) How do these roles relate to Dowty's proto-roles?

* The A0 role corresponds to Dowty's proto-agent role and A1 corresponds to the proto-patient role.

END