# Question classifier for the Question Answering system Hajen

## EDAN50 – Project presentation

# Goals of *Hajen*

- System that can answer questions in a way similar to IBM Watson

- Specifically developed to answer questions on the quiz show Jeopardy!

- Use semantic knowledge extracted from Swedish sources: Wikipedia, encyclopedias, *Göteborgs posten,* others?

- Tribute to Ulf Hannerz, legendary winner of *Kvitt eller dubbelt* in 1957

LUND
UNIVERSITY

# Kvitt eller dubbelt

- *SVT-klassikern* equivalent to *Jeopardy!* in Sweden

# *Kvitt eller dubbelt*: The questions

## På resa i rymden

**250** Vad kallar vi rymdfarare med ett annat namn?
**Astronauter (i Ryssland kosmonauter)**

**500** Vilket land är det hittills enda som skickat upp rymdfärjor?
**USA**

**1000** Vad svävar i rymden och skickar ner tv-program till parabolantennen?
**Satelliter**

**2000** Är det sant att man kan se kinesiska muren med blotta ögat från månen?
**Nej**

**5000** Vilket år landade människan för första gången på månen - 1909, 1969 eller 1999?
**1969**

**10000** På vilken planet har vi landat bilar, och hoppas snart få dit människor?
**Mars**

372

## Handboll

**250** Är en handboll större eller mindre än en fotboll?
**Mindre**

**500** Hur många utespelare per lag finns det på plan?
**Sex**

**1000** Hur många steg får en handbollsspelare springa utan att studsa bollen?
**Tre**

**2000** Vad kan man också kalla ett sjumeterskast?
**Straffkast**

**5000** Vilket svenskt herrlag har vunnit flest SM på senare år - Lugi, Redbergslid eller Skövde?
**Redbergslid**

**10000** Vem har under många år varit framgångsrik handbollstränare för Sverige?
**Bengt** (Bengan) **Johansson**

299

## Aristocats

**250** Vilken sorts djur är huvudpersoner i filmen Aristocats?
**Katter**

**500** Vilken färg har kattungen Marie?
**Hon är vit**

**1000** Hankatten Thomas hjälper Duchess och hennes ungar. Vad heter han i efternamn?
**O´Malley**

**2000** Vad är systrarna Abigail och Amelia och deras Onkel Waldo för sorts fåglar?
**De är gäss**

**5000** Vad heter den elaka betjänten som vill ha bort katterna?
**Edgar**

**10000** Två hundar jagar betjänten. En heter Lafayette. Vad heter den andra, som bestämmer?
**Napoleon**

62

## * Hunden

**250** Vad kallas hundens ungar?
**Valpar**

**500** En arg hund morrar, men vad gör en hund som varnar för fara?
**Skäller**

**1000** Med vilken kroppsdel visar hunden vanligen att den är glad?
**Svansen** (den viftar på svansen)

**2000** Med vilket rovdjur tror man hunden är släkt?
**Vargen**

**5000** Hur ska nosen kännas på en frisk hund?
**Den ska vara fuktig och kall**

**10000** Vad kallas en hundgård för hunduppfödning med ett annat ord?
**Kennel**

149

TEKNIK OCH UPPFINNINGAR

SPORT OCH FRITID

BÖCKER OCH FILM

DJUR OCH NATUR

LUND UNIVERSITY

# Representing the Questions

- We transcribed a set of cards to this format :

  *Number Category Star Subject Points Question Answer*

  *AnswerExplanation CoarseType*

- To have this :

  *149 Animals and Nature * Dog 250 What is a baby dog called ? Puppy*

- Then we transformed into RDF triples.

- Using SPARQL, we can extract easily data from the graph.
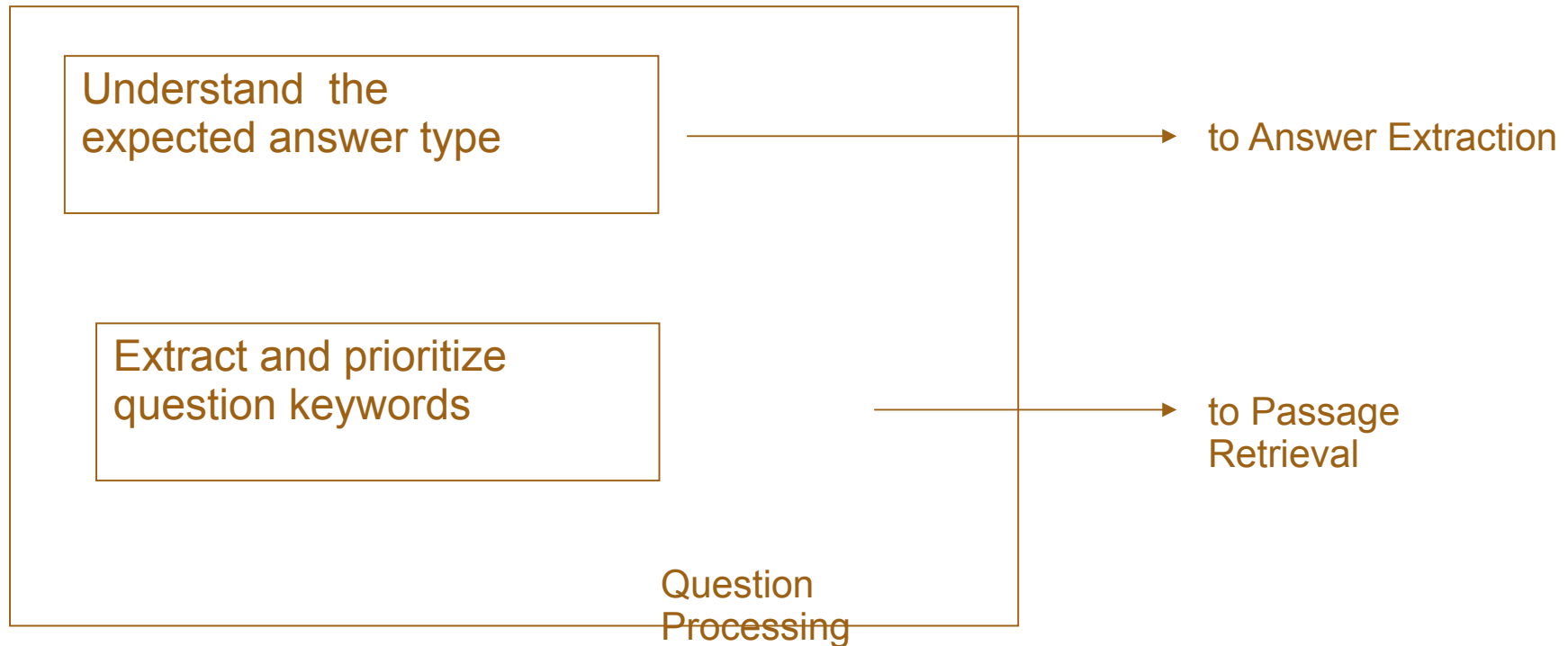
- 230 cards (1380questions) transcribed so far.

# A First Prototype

- Core prototype:

  - A baseline pipeline, two modules so far
    **Question classification** and **Passage retrieval**
    Contributors: Chistopher Käck, Robin Leussier.
    Rebecka Weegar, Juri Pyykkö, students in an AI
    project.

- Additional contributions:

  - **Entity disambiguation**
    Contributor: Marcus Klang, Master's project

  - **Extraction of logical rules from wikipedia**
    Contributors: Alex Wallin and Christian Lindgren, AI
    project

# Question Classification (From Surdeanu and Pasca)

Understand the expected answer type → to Answer Extraction

Extract and prioritize question keywords → to Passage Retrieval

Question Processing

LUND UNIVERSITY

# Question Classification

- Goal: Apply constraints on the answer type

  - Input: A question. *Vad heter den argentinska dans som Petra Nielsen dansade i melodifestivalen 2004?*

  - Output:

    » An answer class. *A description, an entity* (the answer is *Tango*).

    » A question class. *Långt borta och nära*

LUND
UNIVERSITY

# Answer and Question Classes

- Question classes: Category on the back of the card.

- Answer classes:

  - We experimented with classes from *Learning Question Classifiers* by Xin Li and Dan Roth

  - The classes are coarse and fine grained:

    » Course: Entity, description, human, location, numeric.

    » We only worked with the coarse classes

  - We recommend adding to more classes
    » Action and binary

  - Not set in stone. Might need to add classes. Depending on the needs of the passage retrieval

LUND
UNIVERSITY

# Web Interface (API: POST Request)

# Techniques

- Classifiers: LibSVM, LibLINEAR, SMO poly-kernel, BayesNet, SMO StringKernel

- Numerical input vector using the libsvm format except for StringKernel which requires String inputs.

- Vectors represent the presence or absence of a certain word in the question (coverted to lower case and symbols removed)

- Tf-idf weights.
    - WordFreqInQuestion / WordFreqInAllQuestions

LUND
UNIVERSITY

# Training Set

- 1151 questions from *Kvitt eller dubbelt*

- A test set supplied by Quizkampen: 12 categories: *Böcker och ord, Jorden runt*, etc which we used during the development since took long time to transcribe the big dataset.

  - Problem with non-mutually exclusive categories: *Böcker och ord* and *Konst och Kultur* in QK.

  - Not so much in the *Kvitt eller dubbelt* question categories, but in the **answer categories**.

# Results

- Ten fold cross-validation.
- Results on classifying Question classes
  - SVM: Stringkernel          68.9 %
  - SVM: Polykernel            65 %
  - LibSVM: Linear kernel      65.7 %
  - BayesNet                   65 %
  - LinearRegression          68.2 %

  Results from classifying Anser classe
  - SVM: Stringkernel          80.3 %
  - SVM: Polykernel            77.7 %
  - LibSVM: Linear kernel      78.5 %
  - BayesNet                   71.0 %
  - LinearRegression          81.23 %

LUND
UNIVERSITY

# Conclusions

- The StringKernel method performs generally very well, but it is slow to train.
- The improvements in the results in the answer classes seems to be due to a strong bias towards the entity class (most question belongs there).

- Continuing the work it is important to find good answers classes.
  - They need to be useful for the rest of the system
  - They need to be mutually exclusive

- Improving the features
  - We could try stemming, part-of-speech tagging, n-grams etc.

LUND
UNIVERSITY