Probabilistic Domain Model
Elicitation from Natural Language Specifications
work in progress

http://d3s.mff.cuni.cz

Viliam Šimko
simko@d3s.mff.cuni.cz

October 23, 2012
D3S Seminar, Prague
Outline

What
To derive a domain model from text automatically

Why
To simplify initial requirements analysis

How
Supervised Machine Learning
Classification using Maximum Entropy Models
Combining linguistic features and SE artefacts
What would a human analyst do?

- Reading the text
- Sketching the domain model
- Again and again and again ...

Our goal is to automate this
Classical approach

- **Grammatical Inspection** (useful to get a quick start)
  - **Nouns** ~ objects or classes
  - **Adjectives** ~ attributes
  - **Verbs** ~ methods or relationships
  - Removing obviously duplicate terms, e.g.
    - “User Account” = “Customer Account”
    - “Book Review” = “Review Comment”
  - Removing generic and UI-related terms, e.g.
    - “Internet” is too generic
    - “Password” related to UI

Ambiguity, vagueness, uncertainty ...

3.1 User administration

The user administration contains a user account for each user which contains all user data.

A user is able to register at the system with his user number, to manage his user account, and to extend the media’s rental period.

A password is not necessary because the user number on the identification card is read with a bar code scanner.

3.2 Media administration

The media administration contains an entry for each medium in the library. Several instances of each medium may be available, and they may have different locations.

A user is able to search for a medium by specifying one or more features of the media. User can choose and reserve a found instance. To this end, the user number on the identification card is scanned. If a user borrows an instance, it is added to his account and will only be deleted when it is returned.
Ambiguity, vagueness, uncertainty …

3.1 User administration

The user administration contains a user account for each user which contains all user data.

A user is able to register at the system with his user number, to manage his user account, and to extend the media’s rental period.

A password is not necessary because the user number on the identification card is read with a bar code scanner.

3.2 Media administration

The media administration contains an entry for each medium in the library. Several instances of each medium may be available, and they may have different locations.

A user is able to search for a medium by specifying one or more features of the media. User can choose and reserve a found instance. To this end, the user number on the identification card is scanned. If a user borrows an instance, it is added to his account and will only be deleted when it is returned.
Ambiguity, vagueness, uncertainty …

3.1 User administration

The user administration contains a user account for each user which contains all user data.

A user is able to register at the system with his user number, to manage his user account, and to extend the media's rental period.

A password is not necessary because the user number on the identification card is read with a bar code scanner.

3.2 Media administration

The media administration contains an entry for each medium in the library. Several instances of each medium may be available, and they may have different locations.

A user is able to search for a medium by specifying one or more features of the media. User can choose and reserve a found instance. To this end, the user number on the identification card is scanned. If a user borrows an instance, it is added to his account and will only be deleted when it is returned.
To decide, we need training examples...

- **Specification Documents**
- **Domain Model**

Manually annotated text with links to the domain model: "The media administration contains an entry for a medium in the library. Several instances of each medium may be available, and they may have different locations."
The media administration contains an entry for each medium in the library. Several instances of each medium may be available, and they may have different locations.

A librarian is able to add a new media instance to the media administration, change the status of an instance and remove an instance from the media administration.

A user is able to search for a medium by specifying one or more features of the medium. User can choose and reserve a found instance. To this end, the user number on the identification card is scanned. If a user borrows an instance, it is added to his account and will only be deleted when it is returned.
Domain Model of the “Library System”
Linguistic information

Stanford parser: Typed Dependencies, Coreferences, POS, Lemmas, ...

“The media administration contains an entry for each medium in the library.”
Sample Generator

NLP = Extraction of linguistic information
F = Feature Extractor
Established links from annotated text

Derived linguistic information
"User can choose and reserve a found instance"
Features

The word itself
Lemma-form of the word
POS tag of a word
POS tag of a governor word
Type of the dependency between this word and its governor.

etc. …

May be simple paths in the specification model or elaborate mappings.

In our implementation:
Feature ~ Java class implementing the extractor.

\[
\text{e.g. } \text{CurrentWordPosTagExtractor}
\]
\[
\text{w:SpecWord(}
\text{originalWord=“User”}
\text{posTag=“NN”}
\text{)} \rightarrow \text{“pos=NN”}
\]
Thus the experiment ►

Accuracy alone is not sufficient to represent the quality of prediction because the cost of making a FP may be different from the cost of making a FN. F-measure provides a tunable assigned weight for computing a final score and is commonly used to measure the quality of a classification models.

\[
\text{Recall} = \frac{TP}{TP + FN} \\
\text{Precision} = \frac{TP}{TP + FP} \\
F_1 = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}
\]

for \( \alpha = 0.5 \)

\[
\frac{1}{F} = \frac{\alpha}{\text{Recall}} + \frac{1-\alpha}{\text{Precision}}
\]
The Experiment

**Development Phase**

- **Sample Generator**
  - Feature extractors
  - Domain model
  - Specification text
- **Update Specification Model**
  - Specification
- **Generate Samples**
  - Feature sets
  - Samples
- **Evaluation Phase**
  - Performance results
  - $F_1$ of the best model

**Act**

- Split randomly into DEV + TEST
  - Split ratio
  - DEV samples
  - TEST samples

**Feature Sets**

- $\{\text{word, pos, ...}\}$
- $\{\text{pos, indep, ...}\}$
- ...
Development Phase

**Development Phase**

- Try measuring each feature set
- Filtering of features
- **k-fold cross-validation**
- Feature sets

**Aggregate Performance Results**

- input samples
- Split randomly into k subsets
- [i > k]

**MaxEnt Model Evaluation**

- Compute confidence of $F_1$
- Compute $F_1$-measure
- MaxEnt Classification
- Guesses
- Model

---

**Act Development Phase**

- Best feature set

**Act k-fold cross-validation**

- Samples
- K-folds
- Training
- Test subsets

---

**Act MaxEnt Model Evaluation**

- Samples
- Gold contexts
Evaluation Phase

**act Evaluation Phase**

- **train samples**
  - samples
  - feature set

- **feature set**

- **test samples**
  - samples
  - feature set

**MaxEnt Model Evaluation**

- **model**
- **samples**
  - contexts

**MaxEnt Training**

- **Filtered samples**

**Filtered of features**

- **samples**

**MaxEnt Model Evaluation**

- **model**
  - samples
  - filtered samples

**Compute F1-measure**

- **Guesses**
- **gold**

**F1**
Work in progress...

- Implemented
  - Workflow for the experiment (components, ports …)
  - Trainer, classifier, evaluator …
  - Simple Feature Extractors (pos, lemma, indep)
- TODO
  - More Feature Extractors (20+)
  - Sample generator
  - Training data (domain models + annotated texts)
  - Glue code
  - Experimental Results
Thank you for attention