Turning AI Hype into Something Practical: Demystifying Bots

Vincenzo Tursi
Phil Winters

With extensive help from:
Kathrin Melcher
Rosaria Silipo
The History of Bots

1950
A.L.I.C.E.
Artificial Linguistic Internet Computer Entity.

Alan Turing
Bots

“A bot is software designed to automate the kinds of tasks you would usually do on your own (or another human would do for you).”

Search Bots
Teaching Bots
Communication Bots
Personal Assistant Bots
Data & Developer Bots
Team Bots
The Human Internet Search Process

1. Ask Question
2. Translate to Keyword(s)
3. Index
4. Keywords
5. Categories
6. "Best" Question Answer?
   - Yes
   - Not Yet
   - None

© 2018 KNIME AG. All Rights Reserved.
The Search Challenge:  Context

A Jewish Holiday

Paul Shapiro
Professional Search marketer
(and huge KNIME fan…)

Optimizing for Hanukkah: Sometimes it’s still strings, not things

Hannukah
Chanukah
Hanukah
Channukah
Chanuka
Chanukkah
Hanukka
Chanukka
Hannukkah
Hanukkah
Channuka
Festival of Lights
Feast of Dedication
A Data Science Project

Training set

classes

1 2 1 0 4 2 1 0
2 3 1 5 6 2 0 0
6 2 0 0 2 3 1 1
1 5 6 2 3 3 0 3

Data Preparation

Training Model

Apply Model

Scoring

Test set

1 2 1 0 4 2 1 0
2 3 1 5 6 2 0 0
6 2 0 0 2 3 1 1
1 5 6 2 3 3 0 3

Reality Check
Reality Check

classes

Training set

2 1 0  2 1
2 3 1  5 6
6 2  0 2 3 1
5  2 3 0

Test set

2 1 0  2 1
2 1 5 6 2 0
6 2 0 2 3 1
1 5 6 3 3 0

Data Preparation

Training Model

Apply Model

Scoring

We need a class ontology and class labels
Ontology Example

Uberon-an-integrative-multi-species-anatomy-ontology-gb-2012-13-1-r5-2.jpg

Relationship of major animal lineages with indication of how long ago these animals shared a common ancestor. On the left, important organs are shown, which allows us to determine how long ago these may have evolved.
The Specialist Topic Search Process

1. **Ask Question**

2. **Translate to Keyword(s)**
   - **Not Yet**
     - "Best" Question Answer? (Yes/None)

3. **Special Ontology**
   - Keywords: Medicine, symptoms, diseases, treatments
   - Categories: Pharmaceutical, drugs, dosages, allergies
Creating an Ontology: Simple! Build Context

- Ask Question
  - Translate to Keyword(s)
    - Not Yet
      - “Best” Question Answer?
        - Yes
        - None

Keywords
Categories
Ontology
The Dataset

**Questions**

**Community Forum**

2013-2017

**Online Resources**

**E-learning**

~400 Resources

**Blog**

**Other**
Emil the Teacher Bot

- **Ask Question**
- **Understand**
- **Query**
- **Display Resources**
- **Write Feedback**

Keywords:

- Web based user interface
- Text processing NLP functions
- Machine learning model to associate the right group of tutorials to the question that has been asked
- User interface to return the answer
- Feedback mechanism to state whether the answer is of any help
Emil the Teacher Bot: Workflow

This workflow is part of a number of other workflows that address a data mining scenario at the intersection of active learning, text mining, stream mining and service-oriented knowledge discovery architectures. This workflow, in particular, aims to provide a graphical interface for the Web Portal for a KNIME user.

Understanding & Brain

Gather the Question from the User

User Input on the WebPortal

Summary of the Question & Question

Get Probabilities

document class prediction

Choose Answer - Deployment

Switch

Follow Class Path

Probable Predicted Classes

Brain

Concatenate KNIME Resources and Node Guide

Resources

Suggest Resources

Final Class

End Special

No Links Available

End If

End Special

End If

End KNIME Support Team

IF Switch

Switch

Empty Table Switch

Found Links?

Suggest Resources

Have They Benefited from Data and Text Reading

Your Mining Datasets in Optimal Use for Decision-Making

Final Class

End If

End Special

End If

End Special

End KNIME Support Team

End if

Update Datasets

Update Datasets

Update Datasets

Update Datasets
Creation of an Initial Ontology

Ask Question

Teaching bot

Translate to Keyword(s)

Not Yet

"Best" Question Answer?

Yes

None

Email

Keywords

Categories

KNIME Ontology Model

Model Training

Training Set

Initial Labelling
Our Own Ontology (20 Classes)

From e-Learning Course
- Installation
- Data Access
- ETL
- Mining
- Control
- Deployment
- DataViz

From other Resources
- Use Cases
- Text Processing
- Big Data
- Server
- Image Processing
- Reporting

From Experience
- Development
- Integration
- Optimizing KNIME
- Life Science
- Announcement
- Bug
- Legal
Web Crawling KNIME Resources

This workflow reads in the URLs of the different resources available on the KNIME website and extracts the content into a document cell. In case of the node guide only the URL of the first page is given. The metanode extracts all links on the webpage.

Only three nodes
Step 0 – Initial Labeling

Labeling a Training Set based on Distance (and no Clue)
Step 0 – Initial Labeling

**Active Learning 0: Class Labels based on Distance Matching between Resource Documents and Forum Questions**
- Extract keywords from Resource Documents.
- Extract keywords from Forum questions.
- Run a similarity search between keywords vectors from resources and forum questions and find the minimum distance association.
- Assign class to question based on minimum distance association.

**Labeling a Training Set based on Distance (and no Clue)**
Step 1 - Model Training

Train Document Vector and Random Forest Model

This workflow trains two models. A Document Vector Model depending on the keywords of the training set for the PreProcessing and a Random Forest model to make the prediction of the document_class.
Active Learning Cycle

1st attempt
Class Labels

Model Training

Extract most uncertain predictions

Re-labeling

Class Label Extension

Training Set
[Forum Questions]
Active Learning Cycle
Active Learning Cycle

Random Forest

10% most uncertain classes →
Diff. between three top probabilities for each predicted class

Labeling → Predicted Classes or “Something Else”

Labeling manually all “Something Else”

Model Training

Training Set

Initial Labelling

Based on Distance

Active Learning Cycle

k-NN (k=1)

Category Assign

Category Define

Subset chosen to be labeled

Extend

Predicted Classes or “Something Else” Labeling
Step 2a - Category Assign

This workflow is part of a number of other workflows that address a data mining scenario at the intersection of active learning, text mining, stream mining and service-oriented knowledge discovery architectures. This workflow, in particular, provides a graphical interface on the WebPortal for a KNIME specialist to respond to the question with the most uncertain predicted classes. It starts by first reading a subset of the training set (10% of the most uncertain predicted classes). Then, it loops over all the questions, and for each one those, it allows the specialist to choose between one of the predicted classes or the option "Something Else". The labeling phase takes place in the "Choose Answer" webpage. To complete the execution of the loop the specialist has to complete the labeling for all the no-processed questions of the current part of the training set, or to click "Exit". If the specialist clicks "Exit" the workflow saves the last state of the loop iteration. Thus, when the specialist starts again the execution of the workflow on the WebPortal, he/she will be able to start labeling the questions for the last loop iteration. After the Variable Condition Loop End the data is split between questions that haven’t been labeled ("Something Else") and all the other categories. These two datasets are then saved into two different tables.
Adding Active Learning to the Cycle

Ask Question

Teaching bot

Translate to Keyword(s)

Not Yet

No

"Best" Question Answer?

Yes

Email

Keywords

KNIME Ontology Model

Model Training

Training Set

Initial Labelling

Active Learning Cycle

10% lowest probability

Category Assign

Extend

Category Define
Step 3 – Extend with k-NN

- Expert has labelled uncertain samples
- k-NN (k=1) extends the expert classes to their neighbor sample
Step 3 - Extend with k-NN

**Active Learning 5. Extend Oracle Labels to Training Set with k-NN.**
Once we have a subset labelled by the oracle, we need to extend these classes to the whole training set. We do that here with a k Nearest Neighbour.
Adding Active Learning to the Cycle

Ask Question

Teaching bot

Translate to Keyword(s)

"Best" Question Answer?

Not Yet

None

Keywords

Categories

KNIME Ontology Model

Model Training

Training Set

Initial Labelling

10% lowest probability

Active Learning Cycle

Extend

Category Assign

Category Define

Email
Combining the Teaching Bot and the Active Learning Cycle

Teaching bot
- Translate to Keyword(s)
  - Not Yet
  - "Best" Question Answer?
  - Yes
  - None

Ask Question

Keywords
- Categories

KNIME Ontology Model
- Model Training
- Training Set
- Initial Labelling

Active Learning Cycle
- Extend
- Category Assign
- Category Define

10% lowest probability
Changes in Training Set
<table>
<thead>
<tr>
<th>AL #</th>
<th>Input Dataset</th>
<th>Output Model</th>
<th>Ver</th>
<th>Accuracy</th>
<th>Timestamp</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>Training_Set_v0</td>
<td>Random_Forest_v0</td>
<td>0.0</td>
<td>0.59</td>
<td>19/2/2018</td>
</tr>
<tr>
<td>1</td>
<td>Training_Set_v1</td>
<td>Random_Forest_v1</td>
<td>1.0</td>
<td>0.56</td>
<td>23/2/2018</td>
</tr>
<tr>
<td>2</td>
<td>Training_Set_v2</td>
<td>Random_Forest_v2</td>
<td>2.0</td>
<td>0.52</td>
<td>26/2/2018</td>
</tr>
<tr>
<td></td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td></td>
</tr>
</tbody>
</table>

The diagram on the right shows two versions of the KNIME interface, with version 0 and version 2. The table below lists the details of the training set and output model versions.

Here are some links based on your question:
- [KNIME tutorial on using the KNIME server](https://www.knime.com/)
- [KNIME tutorial on creating and deploying a KNIME workflow](https://www.knime.com/)
- [KNIME tutorial on using the KNIME server](https://www.knime.com/)
- [KNIME tutorial on using the KNIME server](https://www.knime.com/)
- [KNIME tutorial on using the KNIME server](https://www.knime.com/)

Question answered? Then click “Yes”. Otherwise click “Back” to choose another category.
Combining the Teaching Bot and the Active Learning Cycle

1. **Ask Question**
   - Teaching bot
     - Translate to Keyword(s)
     - Not Yet
       - "Best" Question Answer?
       - None
       - Yes
         - Email

2. **KNIME Ontology Model**
   - Keywords
   - Categories

3. **Model Training**
   - Training Set
     - Initial Labelling

4. **Active Learning Cycle**
   - Category Assign
   - Category Define

   - 10% lowest probability
   - Extend
Microservices - Converting reusable Subflows into Microservices

Metanode Templates

Microservices
Combining the Teaching Bot and the Active Learning Cycle

Ask Question

Teaching bot

Translate to Keyword(s)

Not Yet

“Best” Question Answer?

None

Email

Keywords

Categories

KNIME Ontology Model

Model Training

Training Set

Initial Labelling

Active Learning Cycle

Extend

Category Assign

Category Define

10% lowest probability
What has been shown ....

• Creating a basic interaction bot
  – Using a workflow for the entire process.
  – Including text mining and categorization

• Building an initial Ontology
  – Using supervised methods and a first “best guess” at categories

• Improving the Ontology with Active Learning
  – Both manual assignment and category creation

• Automating the process
  – Including feedback from users of the bot so that it is “self learning”
How could this be extended?

• Improve text processing phase (tagging)
• Use word embedding
  – Problem: Document Vector leads to big and sparse feature spaces
  – Solution: Train a vector representation for each word using the Word2Vec
• Use the Keras integration to replace the Random Forest with a Neural Network which uses LSTM layers.
• Investigate the role of parameters:
  – 10% of uncertain
  – K=1 in k-Nearest Neighbors
  – Forgetting functions?
• Add speech recognition.
What did we learn?

A bot is always a sequence of 4 activities

- **Interface**
  - text mining
  - Could also include speech recognition as a pre step.
- **Model based on an ontology**
  - And the ontology is critical
  - “accuracy” is not necessarily the best measure of model success!
- **Surface of Relevant results**
  - And feedback on whether those results are relevant.
- **Active Learning to constantly improve model**
  - From Experts
  - From Feedback from users

If you buy a bot:

- Front end is important but don’t let it distract....
- Does the text mining handle your language?
- Does the ontology suit your needs?
- Does the bot self learn and if so, how?
Where to find more

- Presentation available immediately
- Series of blog posts available
- Workflows on EXAMPLE Server