

707.009 Foundations of Knowledge Management „Knowledge Acquisition“

Markus Strohmaier

Univ. Ass. / Assistant Professor
Knowledge Management Institute
Graz University of Technology, Austria

e-mail: markus.strohmaier@tugraz.at
web: <http://www.kmi.tugraz.at/staff/markus>

Overview

- Knowledge Organization
- Broad Knowledge Bases
- Knowledge Acquisition
 - Knowledge and Ontology Engineering
 - Collaborative Knowledge Acquisition
 - Game-Based Knowledge Acquisition



Systems Perspective

acquiring knowledge

Rückblick

Homonyme: Mehrdeutige Benennungen (z.B. Bank)

Homophone: Gleichlautende Benennungen (z.B. Mohr, Moor)

Homographen: Gleiche Schreibweisen (z.B. Wach(-)s(-)tube)

Synonyme: Mehrere Bezeichnungen stehen für denselben Begriff
(Auto, PKW)

Antonyme: Gegensätze (z.B. hart - weich)

Hyper/Hyponyme: Abstraktere / Spezifischere Begriffe (z.B. Fahrzeug / PKW)

Formale Begriffssysteme zielen oft darauf ab **wenig Raum** für Interpretation zu lassen!

- Homonymzusätze (Qualifikatoren)
(z.B. „Ring <Schmuckstück>, Ring <Mathematik>)
- Korrekte Zuordnung von Begriffen und Benennungen oft erst aus dem Kontext heraus interpretierbar!

Retrospect

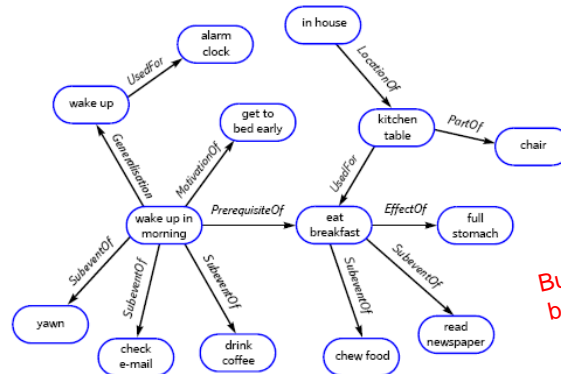
Structure and characteristics of

- Semantic Representations / Ontologies
- WordNet
- ConceptNet
- CyC

Retrospect: ConceptNet

Liu, H. & Singh, P. (2004) ConceptNet: A Practical Commonsense Reasoning Toolkit. BT Technology Journal, Volume 22, Kluwer Academic Publishers.

ConceptNet — a practical commonsense reasoning tool-kit



But how can such broad knowledge bases created?

Fig 1 An excerpt from ConceptNet's semantic network of commonsense knowledge. Compound (as opposed to simple) concepts are represented in semi-structured English by composing a verb (e.g. 'drink') with a noun phrase ('coffee') or a prepositional phrase ('in morning').

Knowledge Acquisition

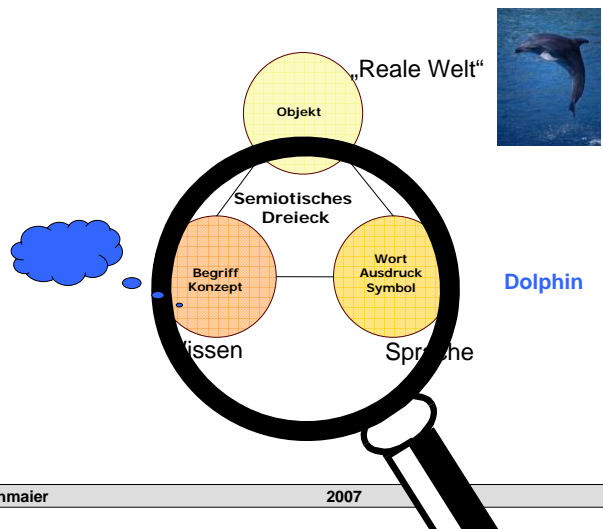
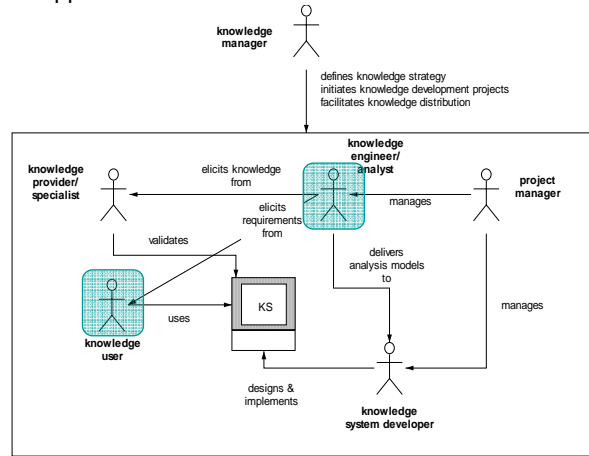
Knowledge acquisition on different levels:

- On an Organizational Level
 - Buying companies, licencing patents, hiring experts
- On an Individual Level
 - Learning, education, training
- On a Technological Level
 - Knowledge bases

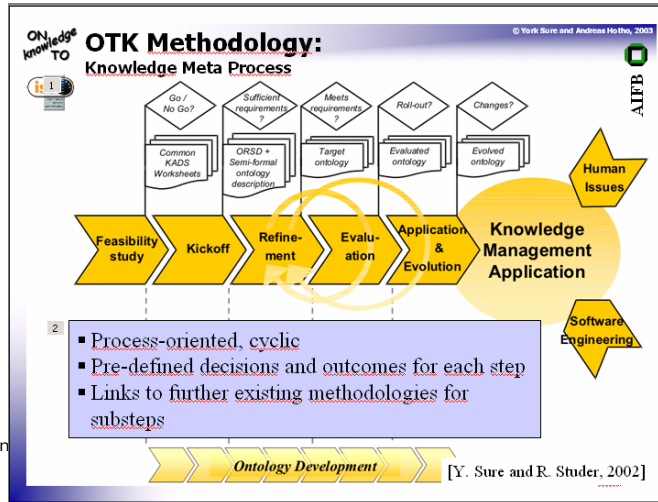
Roles in Knowledge Management and selected KM Support Categories

[Schreiber et al. 02]

The „traditional“ approach:



Der Ontology-Engineering-Prozess [Sure & Studer 2002]



Grafik in Anlehn

Open Mind Common Sense Project <http://commonsense.media.mit.edu/>

Cyc uses paid experts to enter facts in CycL – a proprietary language to represent knowledge

ConceptNet leverages
• User participation

Two types of input:

- Template based acquisition
- Freeform input (restricted in length)

Open Mind Commons
Explain your world.

Home Add new knowledge Highest rated My contributions

Add new knowledge

Example statements

→ water is made of Hydrogen and Oxygen. (by flyboi2003)

→ A credit card is made of plastic. (by Visionsoftkaos)

→ All matter is made of atoms. (by j)

→ a safety pin is made of metal. (by johna)

→ A molecule is made of atoms. (by dbn3)

Teach OpenMind another statement of this type.

is made of .

Teach OpenMind

Open Mind Common Sense Project

<http://commonsense.media.mit.edu/>

Types of relations:

K-LINES (1.25 million assertions) (ConceptuallyRelatedTo 'bad breath' 'mint' 'f=4;i=0;') (ThematicKLine 'wedding dress' 'veil' 'f=9;i=0;') (SuperThematicKLine 'western civilisation' 'civilisation' 'f=0;i=12;')
THINGS (52 000 assertions) (IsA 'horse' 'mammal' 'f=17;i=3;') (PropertyOf 'fire' 'dangerous' 'f=17;i=1;') (PartOf 'butterfly' 'wing' 'f=5;i=1;') MadeOf 'bacon' 'pig' 'f=3;i=0;') (DefinedAs 'meat' 'flesh of animal' 'f=2;i=1;')
AGENTS (104 000 assertions) (CapableOf 'dentist' 'pull tooth' 'f=4;i=0;')
EVENTS (38 000 assertions) (PrerequisiteEventOf 'read letter' 'open envelope' 'f=2;i=0;') (FirstSubeventOf 'start fire' 'light match' 'f=2;i=3;') (SubeventOf 'play sport' 'score goal' 'f=2;i=0;') (LastSubeventOf 'attend classical concert' 'applaud' 'f=2;i=1;')
SPATIAL (36 000 assertions) (LocationOf 'army' 'in war' 'f=3;i=0;')
CAUSAL (17 000 assertions) (EffectOf 'view video' 'entertainment' 'f=2;i=0;') (DesirousEffectOf 'sweat' 'take shower' 'f=3;i=1;')
FUNCTIONAL (115 000 assertions) (UsedFor 'fireplace' 'burn wood' 'f=1;i=2;') (CapableOfReceivingAction 'drink' 'serve' 'f=0;i=14;')
AFFECTIVE (34 000 assertions) (MotivationOf 'play game' 'compete' 'f=3;i=0;') (DesireOf 'person' 'not be depressed' 'f=2;i=0;')

Open Mind Common Sense Project

<http://commonsense.media.mit.edu/>

Liu, H. & Singh, P. (2004) ConceptNet: A Practical Commonsense Reasoning Toolkit. BT Technology Journal, Volume 22, Kluwer Academic Publishers.

Building ConceptNet

1. Extraction phase

50 extraction rules in regular expression form
 Syntactic and semantic constraints are enforced

2. Normalisation phase

Spelling correction, lemmatization (replacing terms with their base form), removal of determiners („the“, „a“)

3. Relaxation phase

Improving the connectivity of the network. Merging duplicate assertions, adding frequency metadata, heuristics. Utilization of WordNet's and FrameNet's synsets and class-hierarchies

Inferring assertions:

```

[ (IsA 'apple' 'fruit');
  (IsA 'banana' 'fruit');
  (IsA 'peach' 'fruit') ]

AND

[ (PropertyOf 'apple' 'sweet');
  (PropertyOf 'banana' 'sweet');
  (PropertyOf 'peach' 'sweet') ]

IMPLIES

(PropertyOf 'fruit' 'sweet')
    
```

Open Mind Common Sense Project

<http://commonsense.media.mit.edu/>

DEMO: Openmind Common Sense
<http://commonsense.media.mit.edu/>

Example:
 A car „is a kind of“ animal



Constructing ConceptNet

<http://commonsense.media.mit.edu/>

Liu, H. & Singh, P. (2004) ConceptNet: A Practical Commonsense Reasoning Toolkit. BT Technology Journal, Volume 22, Kluwer Academic Publishers.

Extraction Phase

Each **node** is an english fragment composed of 4 syntactic constructions:

- Verbs (buy, not eat, drive)
- Noun phrases (red car, laptop computer)
- Prepositional phrases (at work)
- Adjectival phrases (very sour, red)

Verbs must precede noun phrases and adj. Phrases, which in turn must precede prepositional phrases

Illustration:

„If you want to **own an expensive car** then you should **have lots of money or rich parents**“

Constructing ConceptNet

<http://commonsense.media.mit.edu/>

Liu, H. & Singh, P. (2004) ConceptNet: A Practical Commonsense Reasoning Toolkit. BT Technology Journal, Volume 22, Kluwer Academic Publishers.

Normalization Phase

- Unsupervised spellchecker
- Stripping of determiners (the, a)

Lemmatization:

- Words are stripped of tense (is/are/were -> be)
- Plural -> Singular (apples -> apple)

Illustration:

„If you want to own ~~an~~ expensive car then you should have earned lots of money or have rich parents“

Constructing ConceptNet

<http://commonsense.media.mit.edu/>

Liu, H. & Singh, P. (2004) ConceptNet: A Practical Commonsense Reasoning Toolkit. BT Technology Journal, Volume 22, Kluwer Academic Publishers.

Relaxation Phase

Goal: Improve the connectivity of the network

Merge duplicate assertions

Add additional metadata field „frequency“

- *f* counts the number of times a fact is uttered in the OMCS corpus.
- *i* counts how many times an assertion was inferred during the relaxation phase

Produce „intermediate“ knowledge such as semantic and lexical generalisations

Helps bridge other knowledge and to improve the connectivity of the knowledgebase

Apple IsA fruit

„Lifting“ knowledge by leveraging the IsA relationship

```

[ (IsA 'apple' 'fruit');
  (IsA 'banana' 'fruit');
  (IsA 'peach' 'fruit') ]
AND
[ (PropertyOf 'apple' 'sweet');
  (PropertyOf 'banana' 'sweet');
  (PropertyOf 'peach' 'sweet') ]
IMPLIES
(PropertyOf 'fruit' 'sweet')
    
```

„Lifting“ knowledge by leveraging adjectival modifiers

```

[ (IsA 'apple' 'red round object');
  (IsA 'apple' 'red fruit') ]
IMPLIES
(PropertyOf 'apple' 'red')
    
```

Constructing ConceptNet

<http://commonsense.media.mit.edu/>

Liu, H. & Singh, P. (2004) ConceptNet: A Practical Commonsense Reasoning Toolkit. BT Technology Journal, Volume 22, Kluwer Academic Publishers.

Relaxation Phase

Goal: Improve the connectivity of the network

Resolve vocabulary discrepancies and morphological variations (bike / bicycle)

Adding SuperThematicKline to reconcile action/state differences (relax/relaxation) or adjective/nominal differences (sad/sadness)

Utilizing WordNet and FrameNet's verb synonym-sets and class-hierarchies

Using other knowledge bases to aid the knowledge base construction process

Constructing ConceptNet

<http://commonsense.media.mit.edu/>

Liu, H. & Singh, P. (2004) ConceptNet: A Practical Commonsense Reasoning Toolkit. BT Technology Journal, Volume 22, Kluwer Academic Publishers.

Output:

- *f* counts the number of times a fact is uttered in the OMCS corpus.
- *i* counts how many times an assertion was inferred during the relaxation phase

K-LINES (1.25 million assertions)

(ConceptuallyRelatedTo 'bad breath' 'mint' 'f=4;i=0;')
(ThematicKLine 'wedding dress' 'veil' 'f=9;i=0;')
(SuperThematicKLine 'western civilisation' 'civilisation' 'f=0;i=12;')

THINGS (52 000 assertions)

(IsA 'horse' 'mammal' 'f=17;i=3;')
(PropertyOf 'fire' 'dangerous' 'f=17;i=1;')
(PartOf 'butterfly' 'wing' 'f=5;i=1;')
(MadeOf 'bacon' 'pig' 'f=3;i=0;')
(DefinedAs 'meat' 'flesh of animal' 'f=2;i=1;')

AGENTS (104 000 assertions)

(CapableOf 'dentist' 'pull tooth' 'f=4;i=0;')

EVENTS (38 000 assertions)

(PrerequisiteEventOf 'read letter' 'open envelope' 'f=2;i=0;')
(FirstSubeventOf 'start fire' 'light match' 'f=2;i=3;')
(SubeventOf 'play sport' 'score goal' 'f=2;i=0;')
(LastSubeventOf 'attend classical concert' 'applaud' 'f=2;i=1;')

SPATIAL (36 000 assertions)

(LocationOf 'army' 'in war' 'f=3;i=0;')

CAUSAL (17 000 assertions)

(EffectOf 'view video' 'entertainment' 'f=2;i=0;')
(DesirableEffectOf 'sweat' 'take shower' 'f=3;i=1;')

FUNCTIONAL (115 000 assertions)

(UsedFor 'fireplace' 'burn wood' 'f=1;i=2;')
(CapableOfReceivingAction 'drink' 'serve' 'f=0;i=14;')

AFFECTIVE (34 000 assertions)

(MotivationOf 'play game' 'compete' 'f=3;i=0;')
(DesireOf 'person' 'not be depressed' 'f=2;i=0;')

Also see: Wired Magazin Article: "Inside the High Tech Hunt for a Missing Silicon Valley Legend"
http://www.wired.com/techbiz/people/magazine/15-08/ff_jimgray?currentPage=all

Human Computation

Wired News on Steve Fossett gone missing

http://www.wired.com/science/planetearth/news/2007/09/fossett_search_expands



The ARCHER system is attached to a Gippisland GA-8 Airvan like the one shown above.

Photo: Courtesy of the Civil Air Patrol



Here's an example of the size object crowdsource searchers are seeking. The white plane shown above (30-pixel wingspan by 21-pixel length) is approximately the size of Fossett's plane.

Image: Courtesy of Amazon Mechanical Turk

HIT...Human Intelligence Task

Mechanical Turk

The screenshot shows the Amazon Mechanical Turk homepage. At the top, it says "amazonmechanicalturk" and "What is Mechanical Turk?". Below that are navigation tabs for "Welcome", "HITs", and "Qualifications" (with "0,175 HITs available now"). A search bar contains "HITs" and "containing". A yellow banner says "Update on the Steve Fossett search is here". The main text reads: "Complete simple tasks that people do better than computers. And, get paid for it. Learn more." Below this is a section titled "Do you want to quickly and easily create your own HITs on Amazon Mechanical Turk?" with a "Create HITs now" link. The page is divided into three steps:

- STEP 1: Find** - Find HITs to work on. Includes a definition of a HIT and instructions on how to find them.
- STEP 2: Finish** - Work & submit your HIT. Includes instructions on how to complete a HIT.
- STEP 3: Earn** - Get paid for your work. Includes information on how and when you get paid.

 A "Get Started Now" button is at the bottom right.



Here's an example of the size object crowdsource searchers are seeking. The white plane shown above (30-pixel wingspan by 21-pixel length) is approximately the size of Fossett's plane. Image: Courtesy of Amazon Mechanical Turk

Mechanical Turk

Demo: <http://www.mturk.com>

Examples: Powerset, Steve Fossett

The screenshot shows the Amazon Mechanical Turk interface. At the top, it says 'amazonmechanical turk' and 'Artificial Intelligence'. Below that, there are navigation links for 'Welcome', 'MTurks', 'Qualifications', and '15,339 MTurks available now'. A search bar is present with the text 'Search for MTurks containing that pay at least \$ 0.00 for which you are qualified'. There are also buttons for 'Accept MT' and 'Create Your Own MT'. The main task area is titled 'Image Feature Discovery (155696-155699-155695)' with a reward of '\$0.03 per MT' and a duration of '60 minutes'. It includes a list of features to identify: 'Near a Lake / River', 'Near an Interstate/ Near an Exist to the Highway', 'Near Public Transportation (Subway/Bus)', 'In Downtown Area / Center of City', 'Near a Beach / Sea / Ocean', and 'None Of the above'. There are three zoomed-in satellite images of a location labeled 'Location 155696: Zoom 1', 'Zoom 2', and 'Zoom 4'. The zoomed-in images show a street intersection with 'N Rampart Blvd' and 'Rampart Blvd'.

Now, Luis van Ahn, Ass. Prof. at CMU, asks:

What if...

This work would be **fun**?
 And people would **love** to do it?
 Even if they **do not get paid**?

Recommended lecture: <http://video.google.com/videoplay?docid=-8246463980976635143>

Games with a Purpose / Human Computation

EXAMPLE

The Problem:

- Computers are not-so-good in understanding images
- This restricts the ability to provide effective image search engines
- Image search relies on accurate metadata
- **Accurate** metadata is costly to generate

Observation: humans can easily identify complex concepts that consist of multiple parts in images, such as trees, bicycles, ...

Idea: human computation

Games with a Purpose / Human Computation The Google Image Labeler

THE **ESP GAME**

TWO-PLAYER ONLINE GAME

PARTNERS DON'T KNOW EACH OTHER
AND CAN'T COMMUNICATE

OBJECT OF THE GAME:

TYPE THE SAME WORD

THE ONLY THING IN COMMON IS

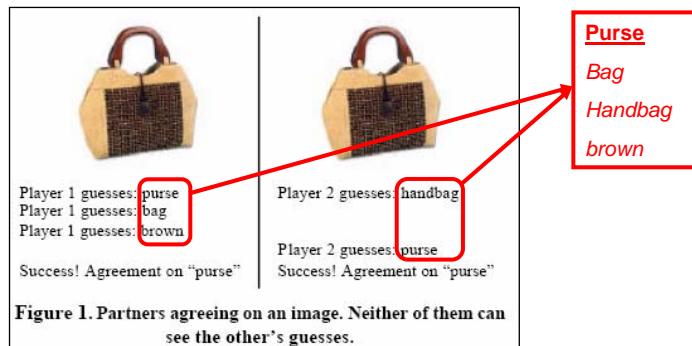
AN IMAGE

Slide: Luis von Ahn

Games with a Purpose / Human Computation The ESP Game

Addresses two problems at the same time:

- 1) metadata **generation** and
- 2) metadata **validation**



L. von Ahn and L. Dabbish. Labeling images with a computer game. In Proceedings of the ACM CHI, 2004.

Games with a Purpose / Human Computation

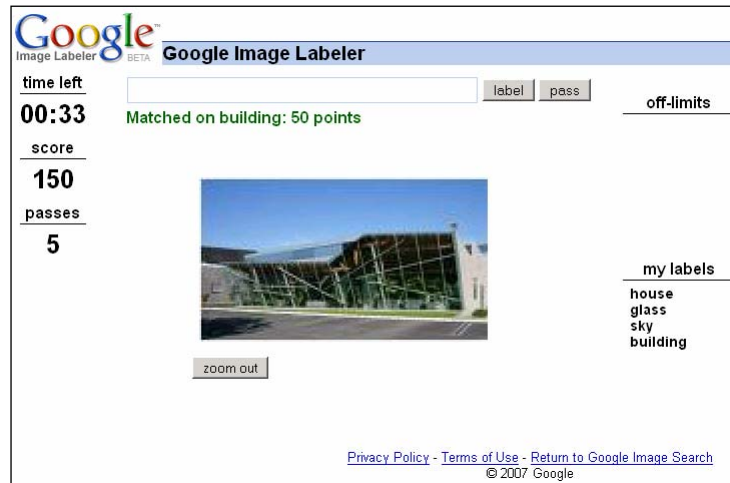
LABELING THE **ENTIRE WEB**

5000 PEOPLE PLAYING SIMULTANEOUSLY CAN LABEL ALL IMAGES ON GOOGLE IN **30 DAYS!**

INDIVIDUAL GAMES IN YAHOO! AND MSN
AVERAGE OVER 5,000 PLAYERS AT A TIME

Slide: Luis von Ahn

The Google Image Labeler



Games with a Purpose / Human Computation

DEMO:

Let's play the Google Image Labeling Game:

<http://images.google.com/imagelabeler/>

*But how can this idea be used
to construct knowledge bases
such as ConceptNet?*

Verbosity

Luis von Ahn, Mihir Kedia and Manuel Blum, Verbosity: a game for collecting common-sense facts, CHI '06: Proceedings of the SIGCHI conference on Human Factors in computing systems, : 75--78, 2006.

Instead of asking users to enter true or false statements, or to rate statements, Verbosity leverages the fact that a game exists that requires users to state common-sense facts: **Taboo™**

Example:

Players have to describe the word „apple“ without saying „apple“ and without saying „red, pie, fruit, macintosh etc.

The player has to give a good enough description of the word to get his teammates guess the right concept („you can make juice out of it“)



The game requires players to say a list of common-sense facts about each word in order to get their teammates to guess it

Verbosity

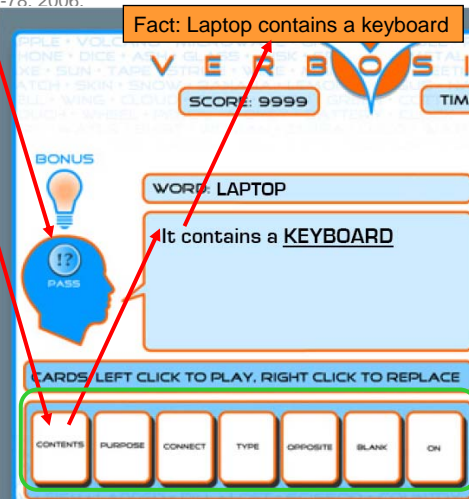
Luis von Ahn, Mihir Kedia and Manuel Blum, Verbosity: a game for collecting common-sense facts, CHI '06: Proceedings of the SIGCHI conference on Human Factors in computing systems, : 75--78, 2006.

One of the players is chosen as the **“Narrator”** while the other is the **“Guesser.”** The Narrator gets a secret word and must get the Guesser to type that word by sending hints to the Guesser.

The hints take the form of **sentence templates** with blanks to be filled in. The Narrator can fill in the blanks with any word they wish except the secret word (or any string containing the secret word).

For example, if the word is LAPTOP, the Narrator might say: “it has a **KEYBOARD.**”

The Guesser guesses. The Narrator can see all of these guesses, and can tell the Guesser whether each is **„hot“** or **„cold“**.



Verbosity

Luis von Ahn, Mihir Kedia and Manuel Blum, Verbosity: a game for collecting common-sense facts, CHI '06: Proceedings of the SIGCHI conference on Human Factors in computing systems, : 75--78, 2006.

Player can describe the secret by using sentence templates **only**.

Reasons:

- Disambiguation
- Categorization
- Parsing
- Fun

Template Examples:

- ___ is a kind of ___
- ___ is used for ___
- ___ is typically near/in/on ___
- ___ is the opposite of ___
- ___ is related to ___
- ___ (wildcard)

Verbosity

Google Techtalks, Luis von Ahn, 2006
<http://video.google.com/videoplay?docid=-8246463980976635143>

NARRATOR

MILK

Is typically near
CEREAL

GUESSER

-> Is typically
near CEREAL

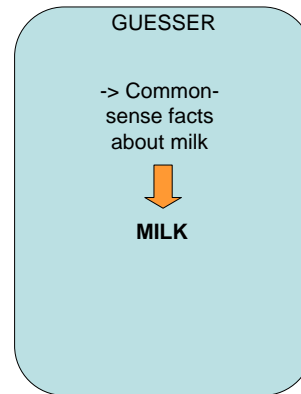
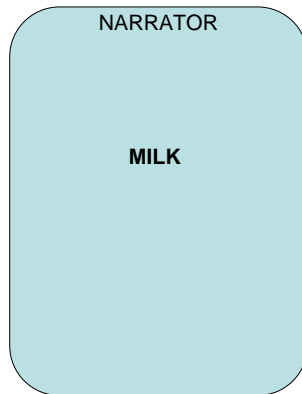
-> Is a LIQUID

-> ...

MILK!

Verbosity

Google Techtalks, Luis von Ahn, 2006
<http://video.google.com/videoplay?docid=-8246463980976635143>



Verbosity

Luis von Ahn, Mihir Kedia and Manuel Blum, Verbosity: a game for collecting common-sense facts, CHI '06: Proceedings of the SIGCHI conference on Human Factors in computing systems, : 75--78, 2006.

Validation and Strategies for assuring accuracy of facts:

- Both Guesser and the Narrator receive points whenever the Guesser enters the correct word
- **Success of the Guesser:** time taken to enter the proper word as an indicator for the quality of the Narrator's statements
- **Random pairing of the players:** Avoiding manipulation
- **Description testing:** Single player mode

Knowledge Management Institute TU
Graz

Verbosity

Google Techtalks, Luis von Ahn, 2006
<http://video.google.com/videoplay?docid=-8246463980976635143>

Asymmetric verification game

Properties: Often fun, verified output

Markus Strohmaier 2007 37

Knowledge Management Institute TU
Graz

Verbosity

Google Techtalks, Luis von Ahn, 2006
<http://video.google.com/videoplay?docid=-8246463980976635143>

Symmetric verification game

If $Output1 = Output2$, both player get points

Markus Strohmaier 2007 38

Verbosity

Luis von Ahn, Mihir Kedia and Manuel Blum, Verbosity: a game for collecting common-sense facts, CHI '06: Proceedings of the SIGCHI conference on Human Factors in computing systems, : 75--78, 2006.

Symmetric vs. Asymmetric games

Symmetric games:

Constraint is number of outputs per input

The ESP game



Asymmetric games:

Constraint is number of inputs that yield the same output

Verbosity

-> Is typically near CEREAL

-> Is a LIQUID

Any questions?

See you next week!