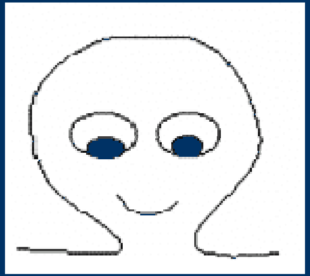
Neurocognitive Inspirations in Natural Language Processing



Włodzisław Duch & Co.

Department of Informatics,
Nicolaus Copernicus University, Poland

School of Computer Engineering,
Nanyang Technological University, Singapore

Google: Duch

Plan

Goal: Reaching human-level competence in all aspects of NLP.

- Neurocognitive inspirations: how are words represented in brains?
- Priming, brains and creativity.
- Morphological level – creating novel words.
- Semantic memory and other types of memory.
- Taking heads and words games.
- A priori knowledge in document categorization, or how to capture intuition in a simple model.
- Enhancing document representations using ontologies and semantic memories.

- Few conclusions while we are still running.

Ambitious approaches…

CYC, Douglas Lenat, started in 1984.
Developed by CyCorp, with 2.5 millions of assertions linking over 150.000 concepts and using thousands of micro-theories (2004).

Cyc-NL is still a “potential application”, knowledge representation in frames is quite complicated and thus difficult to use.

Open Mind Common Sense Project (MIT):

a WWW collaboration with over 14,000 authors, who contributed 710,000 sentences; used to generate ConceptNet, very large semantic network.

Some interesting projects are being developed now around this network but no systematic knowledge has been collected

Other such projects: HowNet (Chinese Academy of Scienc FrameNet (Berkley), various large-scale ontologies.

The focus of these projects is to understand all relations in text/dialogue

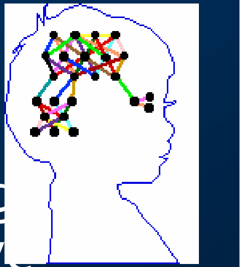
Neurocognitive approach

Why is NLP so hard? Only human brains are adapted to it.

Ambitious approach: make an artificial brain!

Computational cognitive neuroscience: aims at rather detailed neural models of cognitive functions, first annual CNN conf. Nov. 2005.

Brain simulation with ~10^{10} neurons and ~10^{15} synapses (NSI San Diego),
1 sec = 50 days on a 27 processor Beowulf cluster.

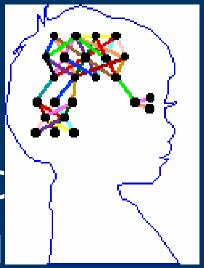


Neurocognitive informatics: focus on simplified models of hig cognitive functions: in case of NLP various types of associative memory: recognition, semantic and episodic.

Many speculations, because we do not know the underlying brain processes, but models explaining most neuropsychological syndromes exist; computational psychiatry is rapidly developing since ~ 1995.

“Roadmap to human level intelligence” – workshops ICANN’05, WCCI’06

Words in the brain



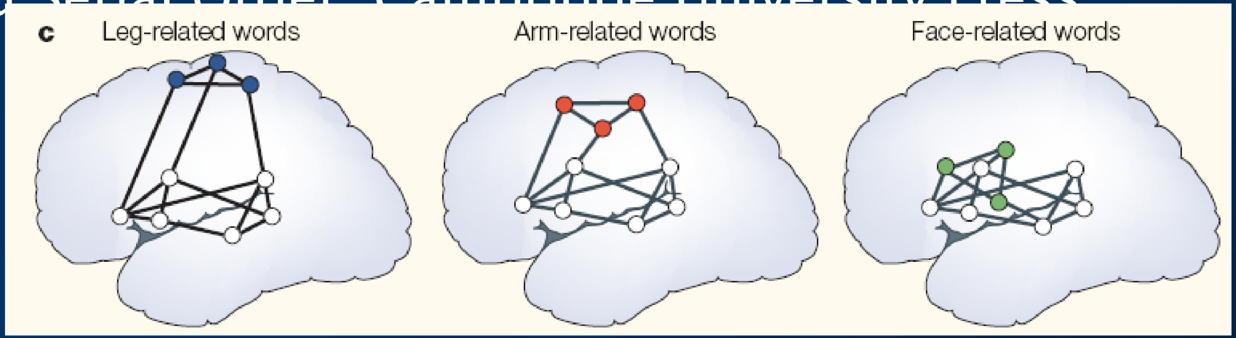
Psycholinguistic experiments show that most likely categoric phonological representations are used, not the acoustic inpu

Acoustic signal => phoneme => words => semantic concepts.

Phonological processing precedes semantic by 90 ms (from N200 ERPs).

F. Pulvermuller (2003) The Neuroscience of Language. On Brain Circuits of Words and Serial Order. Cambridge University Press.

Action-perception networks inferred from ERP and fMRI



Phonological neighborhood density = the number of words that are similar in sound to a target word. Similar = similar pattern of brain activations.

Semantic neighborhood density = the number of words that are similar in meaning to a target word.

Brain areas involved

Organization of the word recognition circuits in the left temporal lobe has been elucidated using fMRI experiments (Cohen et al. 2004).

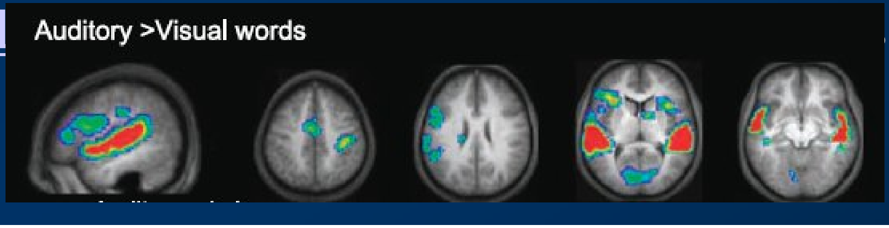
How do words that we hear, see and thinking of activate the brain?

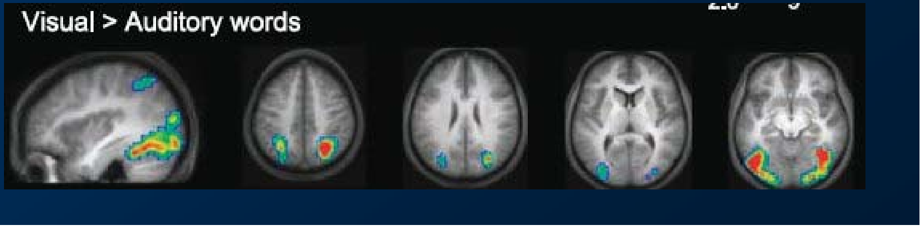
Seeing words: orthography, phonology, articulation, semantics.

Visual word form area (VWFA) in the left occipitotemporal sulcus is strictly unimodal visual area.

Adjacent lateral inferotemporal multimodal area (LIMA) reacts to both auditory & visual stimulation, has cross-modal phonemic and lexical links.

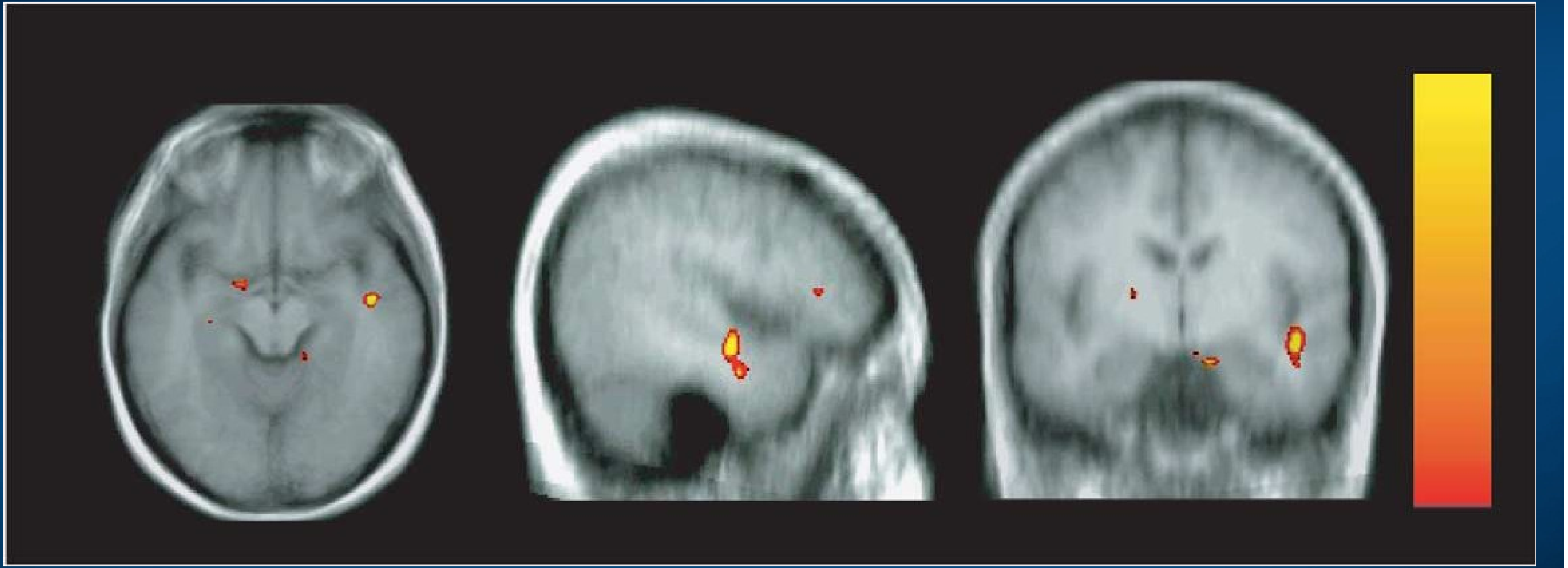
Likely: homolog of the VWFA in the auditory stream, the auditory word form area, located in the left anterior superior temporal sulcus; this area shows reduced activity in developmental dyslexics.





Insights and brains

Activity of the brain while solving problems that required insight and that could be solved in schematic, sequential way has been

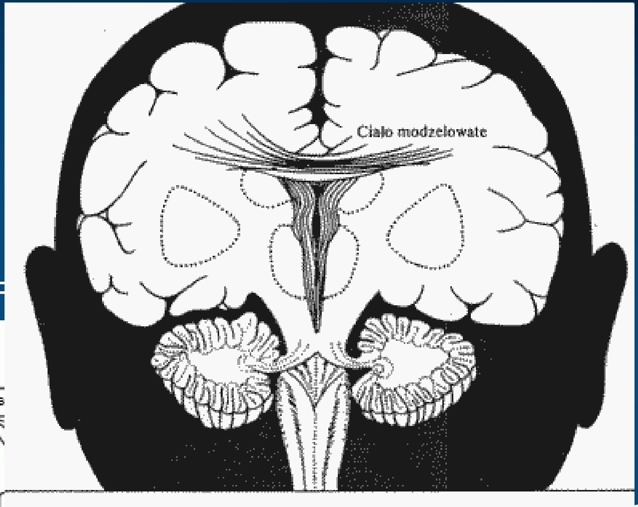


An increased activity of the right hemisphere anterior superior temporal gyrus (RH-aSTG) was observed during initial solving efforts and insights. About 300 ms before insight a burst of gamma activity was observed, interpreted by the authors as „making connections across distantly related information during comprehension ... that allow them to see connections that previously eluded them”.

Insight interpreted

What really happens? My interpretation:

- LH-STG represents concepts, S=Start, F=
- unders
- if no c
- RH-ST … cepts into abstra
- conne … eeling of vague
- gamm … s for S, F and in
- stepwi
- finding … xperience; they a … nanent links.



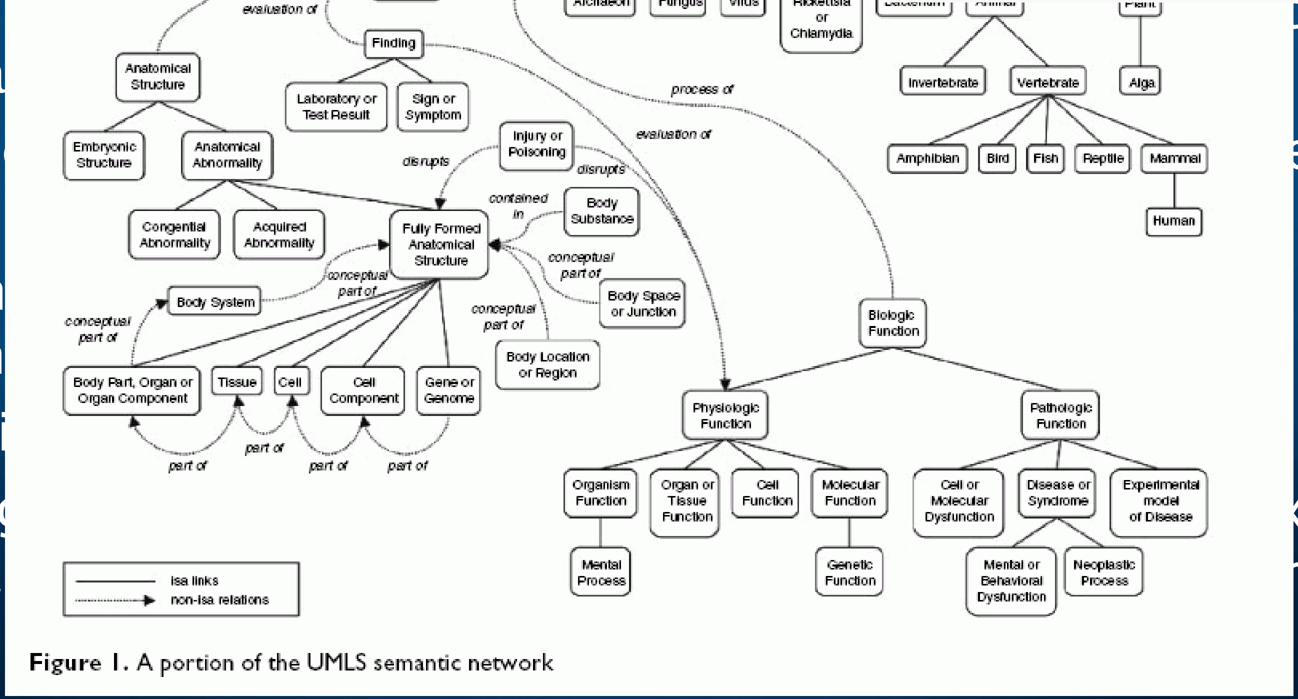


Figure 1. A portion of the UMLS semantic network

Creativity

What features of our brain/minds are most mysterious?

Consciousness? Imagination? Intuition? Emotions, feelings?
Higher mental functions?

Masao Ito (director of RIKEN, neuroscientist) answered: creativity.

Still domain of philosophers, educators and a few psychologists, for ex. Eysenck, Weisberg, or Sternberg (1999), who defined creativity as:
"the capacity to create a solution that is both novel and appropriate".

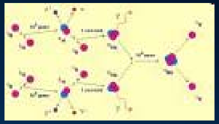
MIT Encyclopedia of Cognitive Sciences has 1100 pages, 6 chapters about logics & over 100 references to logics in the index.

Creativity: 1 page (+1 page about „creative person").

Intuition: 0, not even mentioned in the index.
In everyday life we use intuition more often than logics.

Unrestricted fantasy? Creativity may arise from higher-order schemes!

Memory & creativity



Creative brains accept more incoming stimuli from the surrounding environment (Carson 2003), with low levels of latent inhibition responsible for filtering stimuli that were irrelevant in the past.

“Zen mind, beginners mind” (S. Suzuki) – learn to avoid habituation!

Creative mind maintains complex representation of objects and situations.

Pair-wise word association technique may be used to probe if a connection between different configurations representing concepts in the brain exists.

A. Gruszka, E. Nęcka, Creativity Research Journal, 2002.

Word 1 Priming 0,2 s Word 2

Words may be close (easy) or distant (difficult) to connect;
priming words may be helpful or neutral;
helpful words are related semantically or phonologically (hogse for horse);
neutral words may be nonsensical or just not related to the presented pair.

Creativity & associations

Hypothesis: creativity depends on the associative memory, ability to connect distant concepts together

Results: creativity is correlated with greater & susceptibility to priming, distal associatic before decision is made.

Neutral priming is strange!

- for close words and nonsensical priming worse
  than less creative; in all other cases they c

- for distant words priming always increase association,
  the effect is strongest for creative people.

Latency times follow this strange patterns.

Conclusions of the authors:
More synaptic connections => better assoc creativity.

But results for neutral priming are puzzling!

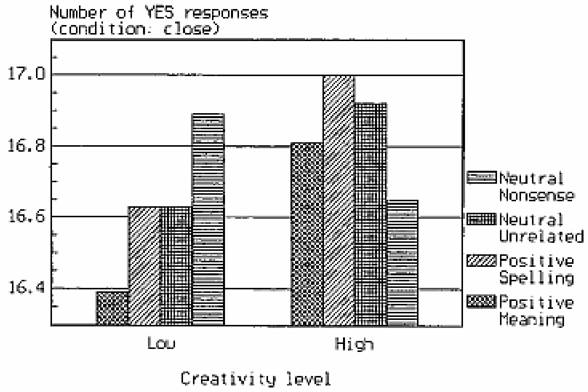


Figure 3. *The acceptance of close associative connections depending on the type of priming and the level of creativity.*

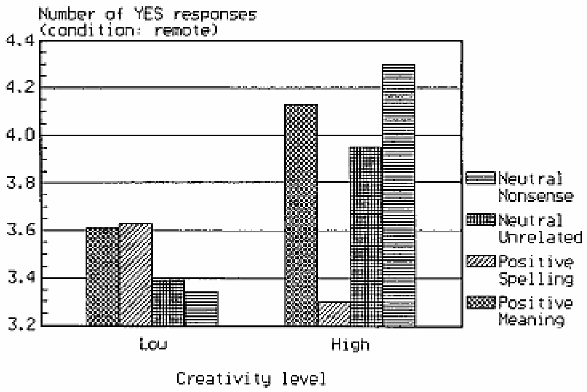
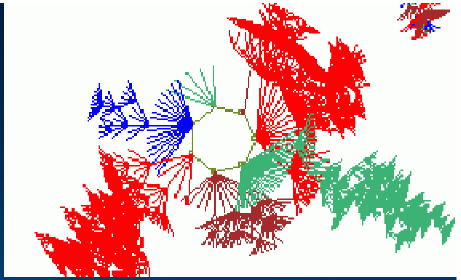
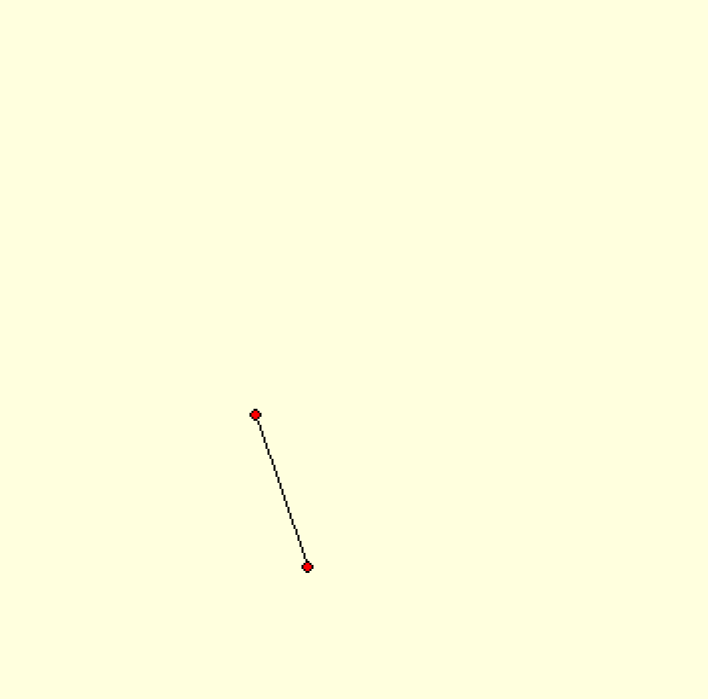


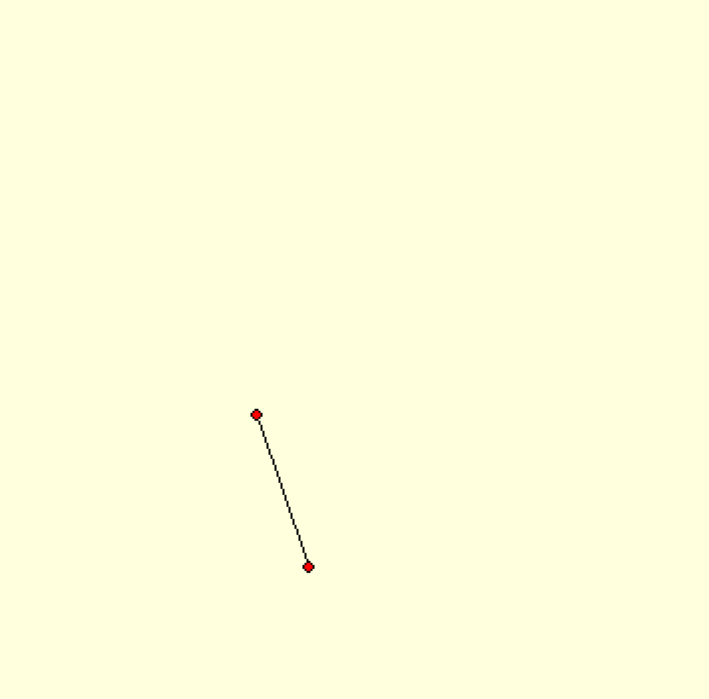
Figure 4. *The acceptance of remote associative connections depending on the type of priming and the level of creativity.*

Paired associations



So why neutral priming for close associations and nonsensical priming words degrades results of creative people?





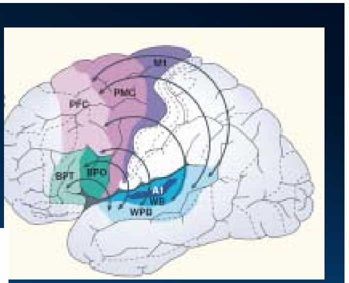
uits;
en many
creates
use the

not help,
contribute

mediate
n distant
hance,
observed in perception.

For priming words with similar spelling and close words the activity of the second word representation is higher, always increasing the chance of connections and decreasing latency. For distant words it

Words: simple model



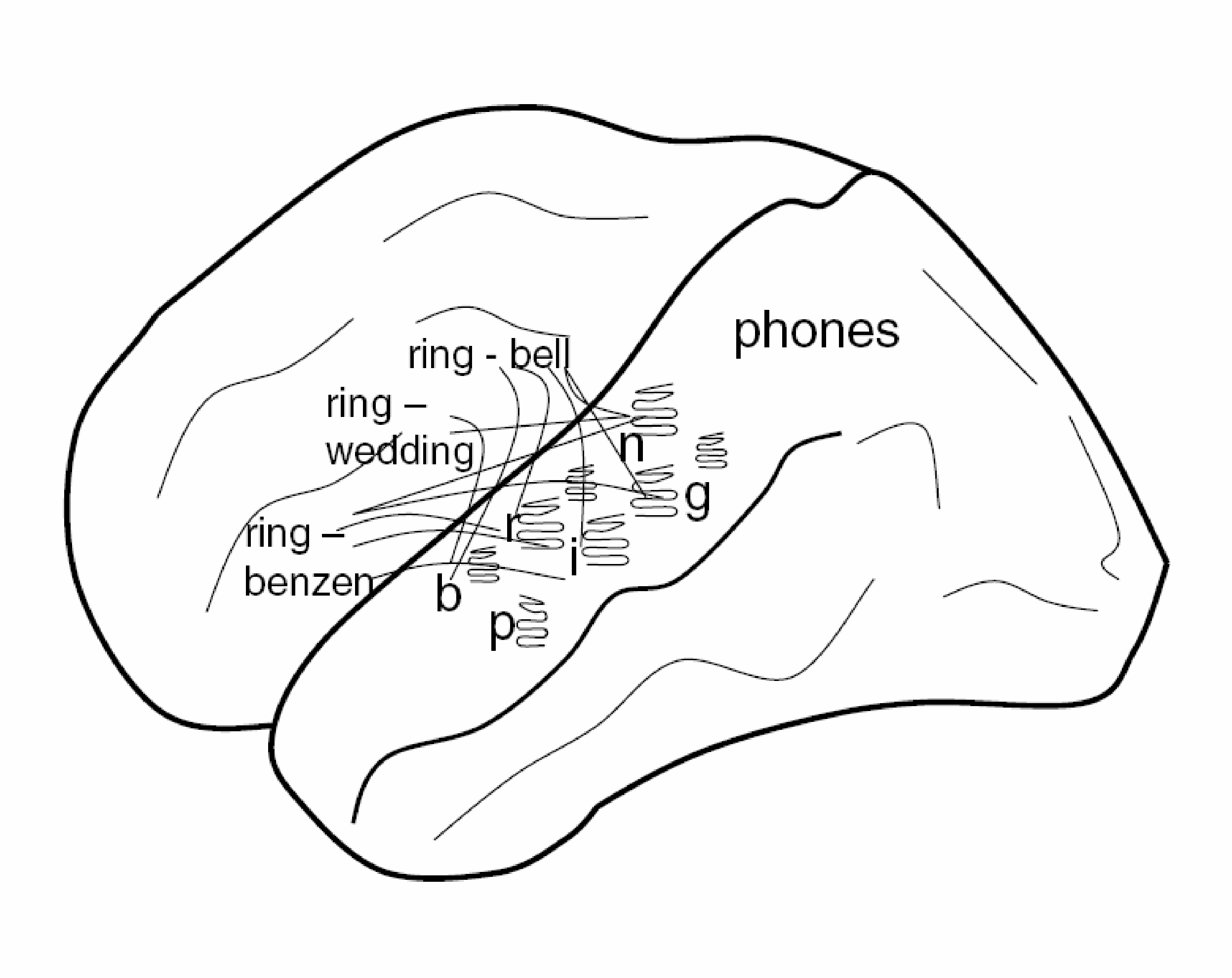
Goals:

- make t
- create i ... es of products;
- underst ... ionary.

Model insp ... words are being inve ... tory cortex.

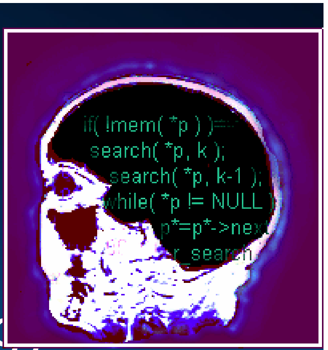
Phonemes ... of phonemes ... combinati ... ost leaves only a few

Creativ ... filtering



Imagination: chains of phonemes activate both word and non-word representations, depending on the strength of the synaptic connections. **Filtering**: based on associations, emotions,

Generating novel words



```
if( !mem( *p ) )==
 search( *p, k );
   search( *p, k-1 );
  while( *p != NULL )
        p*=p*->next
          r_search
```

Approximations: associative neural networks, self-organizi networks, or statistical models capturing phono/morpholog

Preliminary:

- create probability models for linking phonemes and syllables;
- create semantic and phonological distance measures for words.

Statistical algorithm to find novel words:

- Read initial pool of keywords.
- Find phonological and semantic associations to increase the pool.
- Break all words into chains of phonemes and chains of morphemes.
- Find all combinations of fragments forming longer chunks ranked according to their phonological probability (estimating ngram plausibility).
- For final ranking use estimation of semantic density around

Words: experiments

A real letter from a friend:

I am looking for a word that would capture the following qualities: portal to new worlds of imagination and creativity, a place where visitors embark on a journey discovering their inner selves, awakening the Peter Pan within.
A place where we can travel through time and space (from the origin to the future and back), so, its about time, about space, infinite possibilities.
FAST!!! I need it sooooooooooooooooooooooon!

creativital, creatival (creativity, portal), used in creatival.com
creativery (creativity, discovery), creativery.com
(strategy+creativity)
discoverity = {disc, disco, discover, verity} (discovery, creativity, verity)
digventure ={dig, digital, venture, adventure, venue, nature} still new!
imativity (imagination, creativity); infinitime (infinitive, time)
infinition (infinitive, imagination), already a company name
learnativity (taken, see http://www.learnativity.com)
portravel (portal, travel); sportal (space, sport, portal), taken
quelion – lion of query systems! Web site

Word games

Word games were popular before computer games.
They are essential to the development of analytical thinking.
Until recently computers could not play such games.

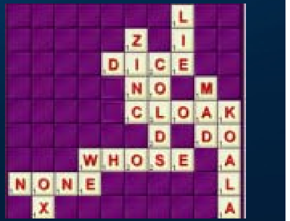
The 20 question game may be the next great challenge for AI, because it is more realistic than the unrestricted Turing test; a World Championship could involve human and software players.

Finding most informative questions requires knowledge and creativity.

Performance of various models of semantic memory and episodic memory may be tested in this game in a realistic, difficult application.

Asking questions to understand precisely what the user has in mind is critical for search engines and many other applications.

Creating large-scale semantic memory is a great challenge: ontologies, dictionaries (Wordnet), encyclopedias, MindNet (Microsoft), collaborative projects like Concept Net (MIT) …



Realistic goals?

Different applications may require different knowledge representation.

Start from the simplest knowledge representation for semantic memory.

Find where such representation is sufficient, understand limitations.

Drawing on such semantic memory an avatar may formu
may answer many questions that would require exponen
large number of templates in AIML or other such languag

Adding intelligence to avatars involves two major tasks:

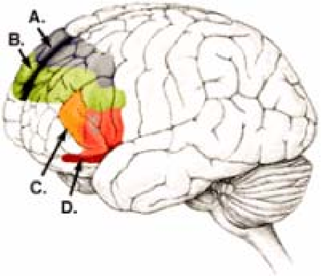
- building semantic memory model;
- provide interface for natural communication.

Goal:

create 3D human head model, with speech synthesis & recognition, use it to interact with Web pages & local programs: a Humanized InTerface (HIT).

Control HIT actions using the knowledge from its semantic memory.

Types of memory



Neurocognitive approach to NLP: at least 4 types of memories:
Long term (LTM): recognition, semantic, episodic + working memory.

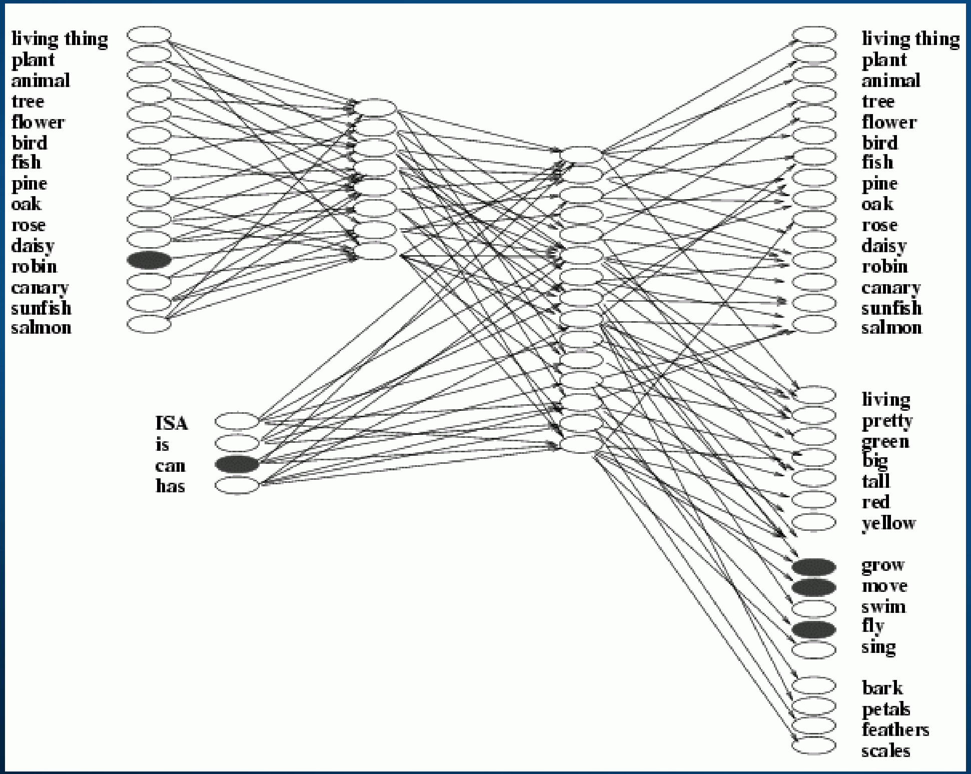
Input (text, speech) pre-processed using recognition memory model to correct spelling errors, expand acronyms etc.
For dialogue/text understanding episodic memory models are needed.
Working memory: an active subset of semantic/episodic memory.
All 3 LTM are coupled mutually providing context for recogniton.

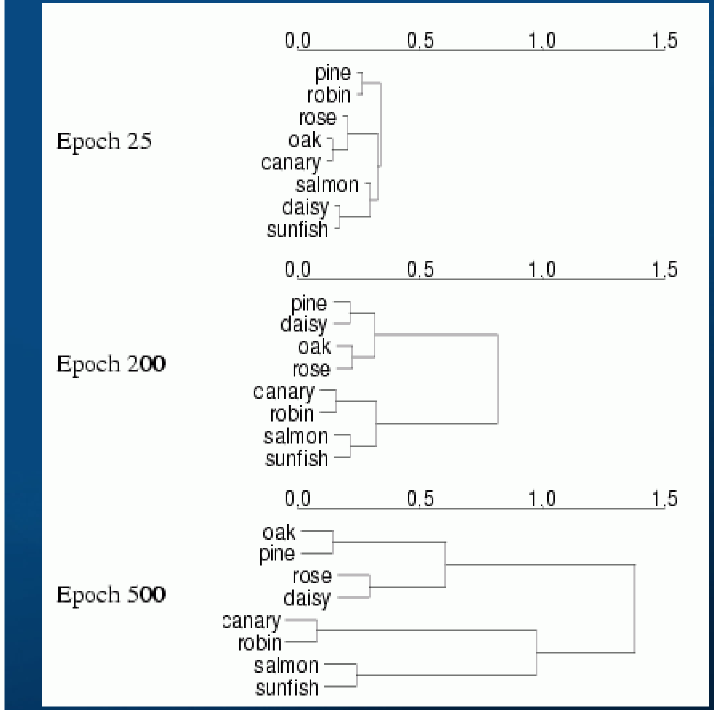
Semantic memory is a permanent storage of conceptual data.

- "Permanent": data is collected throughout the whole lifetime of the system, old information is overridden/corrected by newer input.
- "Conceptual": contains semantic relations between words and uses them to create concept definitions.

SM & neural distances

Activations of groups of neurons presented in activation space define similarity relations in geometrical model.

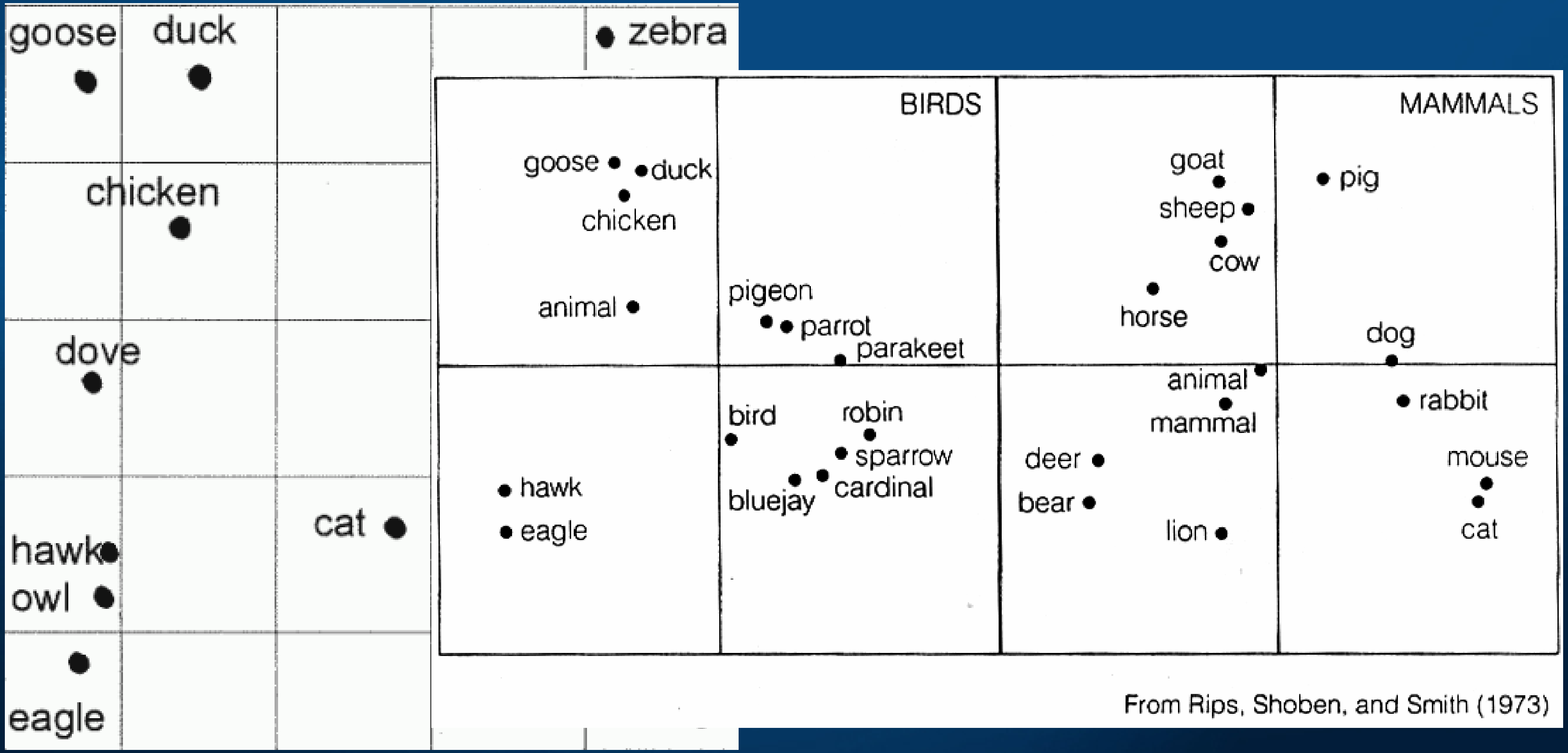




Similarity between concepts

Left: MDS on vectors from neural network.
Right: MDS on data from psychological experiments with perceived similarity between animals.

Vector and probabilistic models are approximations to this process.



From Rips, Shoben, and Smith (1973)

Semantic memory

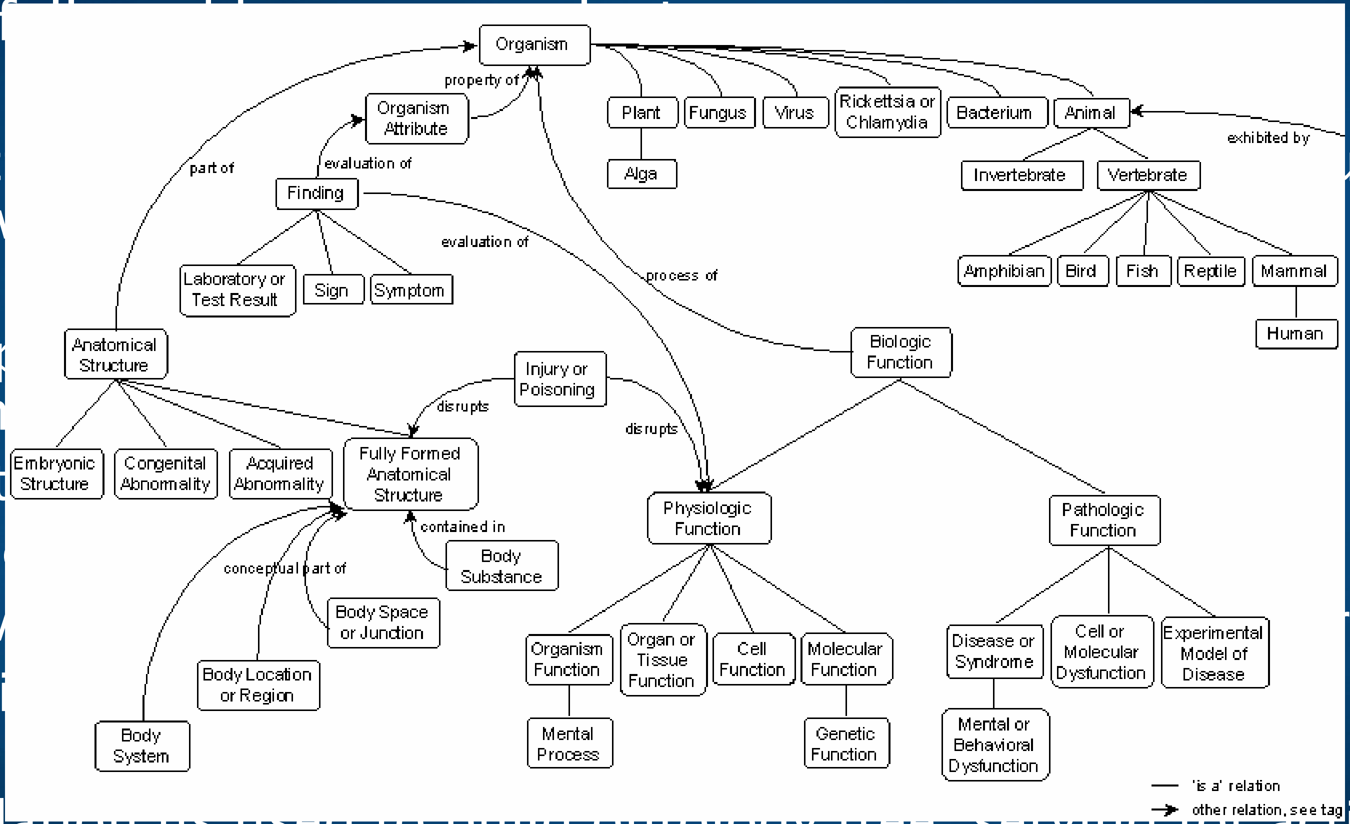
Hierarchical model of semantic memory (Collins and Quillian, 1969), f

Connec us, 1975), w

Our imp relation

The dat

- conc
- keyw rces);
- relati



IS-A relation us used to build ontology tree, serving for activation spreading, i.e. features inheritance down the ontology tree.

Types of relations (like “x IS y”, or “x CAN DO y” etc.) may be defined when input data is read from dictionaries and ontologies.

Creating SM

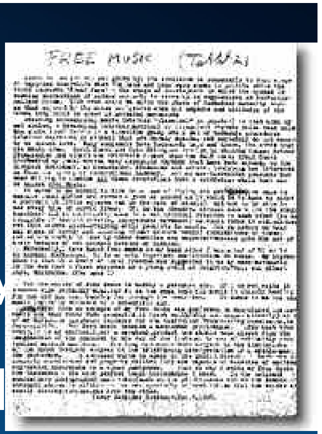
The API serves as a data access layer providing logical operations between raw data and higher application lay Data stored in the database is mapped into application objects and the API allows for retrieving specific concepts/keywords.
Two major types of data sources for semantic memory:

1. machine-readable structured dictionaries directly convertible into
   semantic memory data structures;
2. blocks of text, definitions of concepts from dictionaries/encyclopedias.

3 machine-readable data sources are used:

- The Suggested Upper Merged Ontology (SUMO) and the the MId-Level Ontology (MILO), over 20,000 terms and 60,000 axioms.
- WordNet lexicon, more than 200,000 words-sense pairs.
- ConceptNet, concise knowledgebase with 200,000 assertions.

Creating SM – free text



WordNet hypernymic (a kind of …) IS-A relation + Hypony and meronym relations between synsets (converted into concept/concept relations), combined with ConceptNet rel such as: CapableOf, PropertyOf, PartOf, MadeOf ...

Relations added only if in both Wordnet and Conceptnet.

Free-text data: Merriam-Webster, WordNet and Tiscali.

Whole word definitions are stored in SM linked to concepts.

A set of most characteristic words from definitions of a given concept.

For each concept definition, one set of words for each source dictionary is used, replaced with synset words, subset common to all 3 mapped back to synsets – these are most likely related to the initial concept.

They were stored as a separate relation type.

Articles and prepositions: removed using manually created stop-word list.

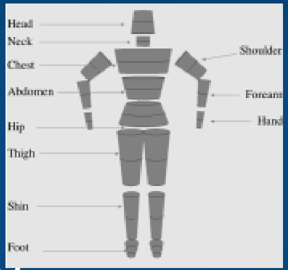
Phrases were extracted using ApplePieParser + concept-phrase relations compared with concept-keyword, only phrases that

Concept Description Vectors

Drastic simplification: for some applications SM is used in a more efficient way using vector-based knowledge representation.

Merging all types of relations => the most general one:

“x IS RELATED TO y”, defining vector (semantic) space.



{Concept, relations} => Concept Description Vector, CDV.

Binary vector, shows which properties are related or have sense for a given concept (not the same as context vector).

Semantic memory => CDV matrix, very sparse, easy storage of large amounts of semantic data.

Search engines: {keywords} => concept descriptions (Web pages).

CDV enable efficient implementation of **reversed queries**:

find a unique subsets of properties for a given concept
or a class of concepts = concept higher in ontology.

What are the unique features of a sparrow? Proteoglycan? Neutrino?

HIT the Web

Haptek avatar as a plug-in in WWW browser.

Connect to web pages, read their contents,
send queries and read answers from specific fields in web forms.

Access Q/A pages, like MIT Start, or Brainboost that answer reasonably to many questions.

“The HAL Nursery”, “the world's first Child-Machine Nursery”, Ai Research www.a-i.com, is hosting a collection of “Virtual Children”, or HAL personalities developed by many users through conversation.

HAL is using reinforcement learning techniques to acquire language, through trial and error process similar to that infants are using.

A child head with child voice makes it much more interesting to play with.

Haptek heads may work with many chatterbots, we focus on use of

Talking Head

SM is the brain, HIT needs a talking head and voice interface.

Haptek's PeoplePutty tools have been used (inexpensive) to creat a 3-D talking head; only the simplest version is used.

Haptek player is a plugin for Windows browsers, or embedded component in custom programs; both versions were used.

High-fidelity natural voice synthesis with lips synchronization may be added to Haptek characters.

Free MS Speech Engine, i.e. MS Speech API (SAPI 5) has been used to add text to speech synthesis and speech to text voice recognition.

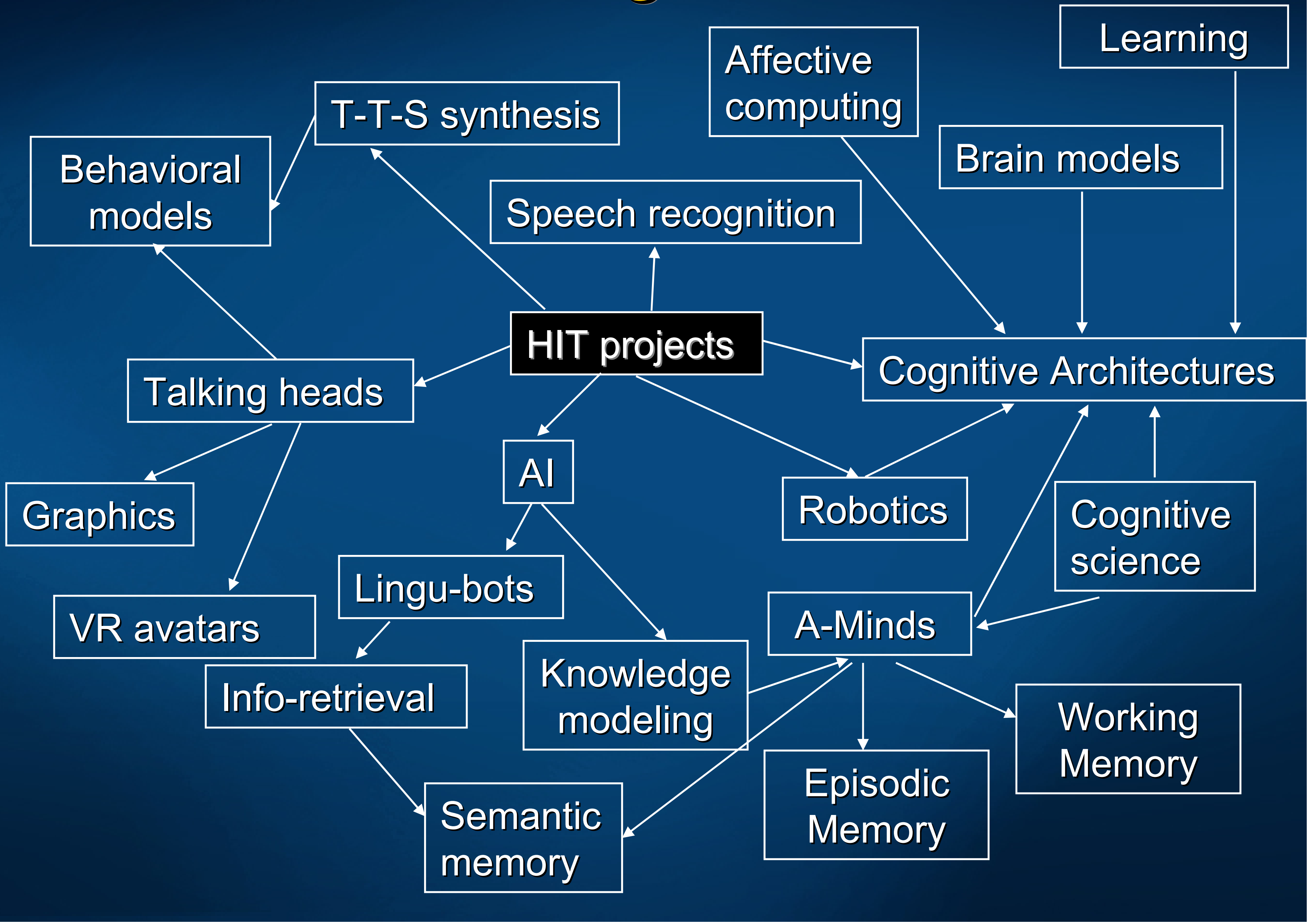
OGG prerecorded audio files may be played.

Haptek movements, gestures, face expressions and animation sequences may be programmed and coordinated with speech using JavaScript, Visual Basic, Active-X Controls, C++, or ToolBook.

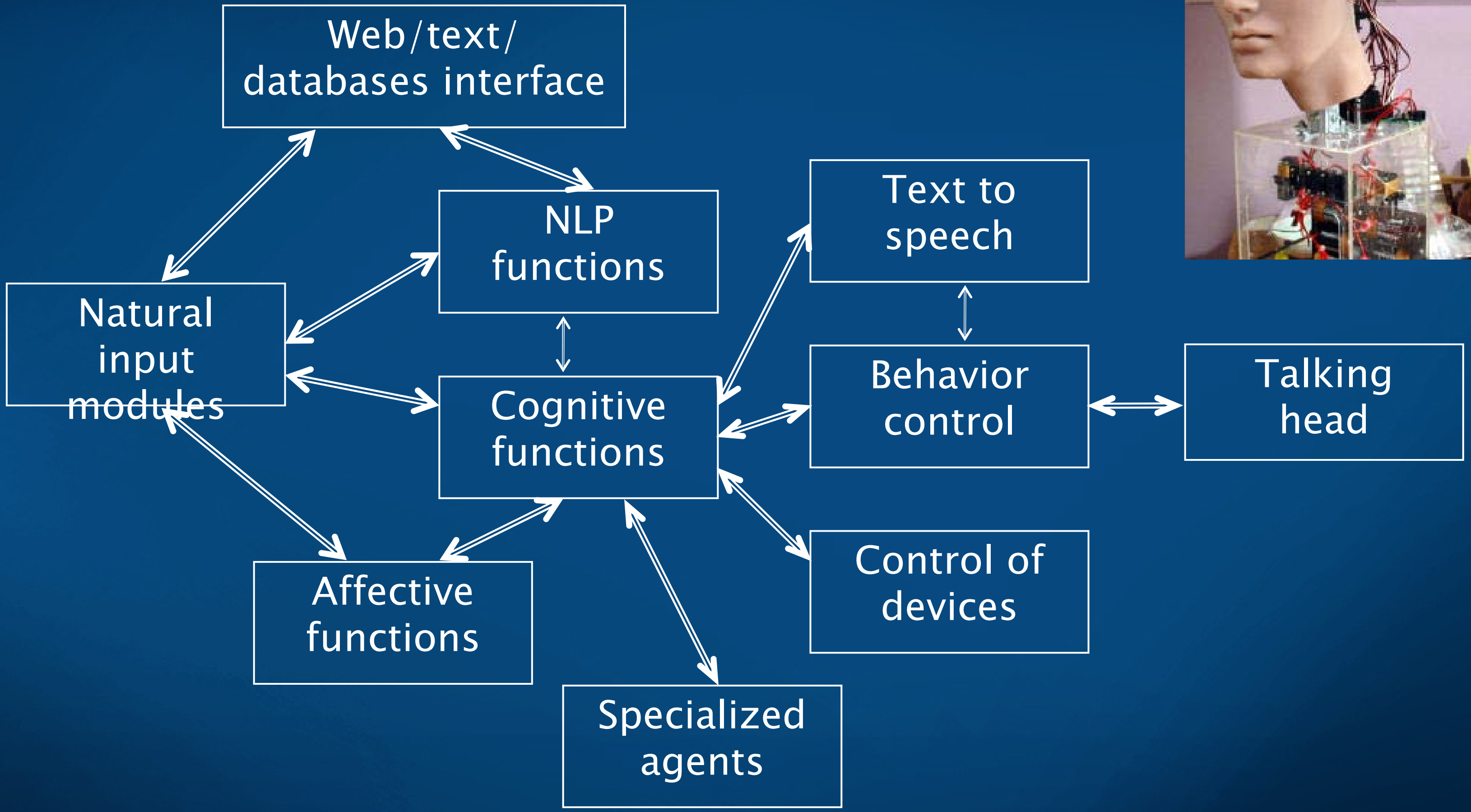
Result: HIT that can interact with web pages, listen and talk, sending information both ways, hiding the text pages from the user.

Interaction with Web pages is based on Microsoft .NET framework

HIT – larger view …

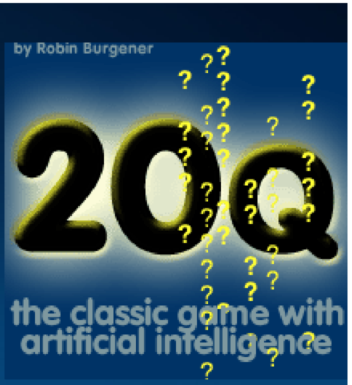


DREAM architecture



DREAM is concentrated on the cognitive functions + real time control, we plan to adopt software from the HIT project for perception, NLP, and other functions.

20Q



The goal of the 20 question game is to guess a concept that the opponent has in mind by asking appropriate questions.

www.20q.net has a version that is now implemented in some toys!
Based on concepts x question table T(C,Q) = usefulness of Q for C.
Learns T(C,Q) values, increasing after successful games, decreasing after lost games. Guess: distance-based.

SM does not assume fixed questions.
Use of CDV admits only simplest form “Is it related to X?”, or “Can it be associated with X?”, where X = concept stored in the SM.
Needs only to select a concept, not to build the whole question.

Once the keyword has been selected it is possible to use the full power of semantic memory to analyze the type of relations and ask more sophisticated questions.
How is the concept selected?

Distance calculation

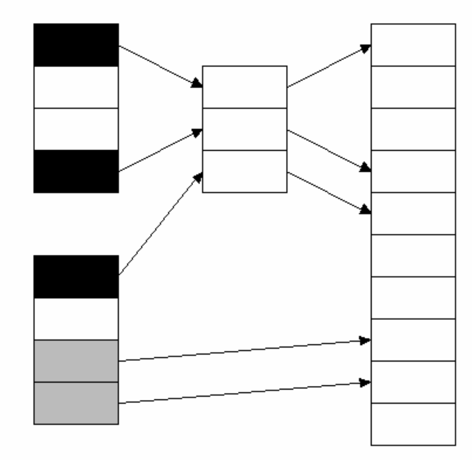
Euclidean distance used for binary Yes/No answer, otherwise the distance ||K–A|| is:

$$\|K - A\| = \sqrt{\sum_i |K_i - A_i|^2}$$

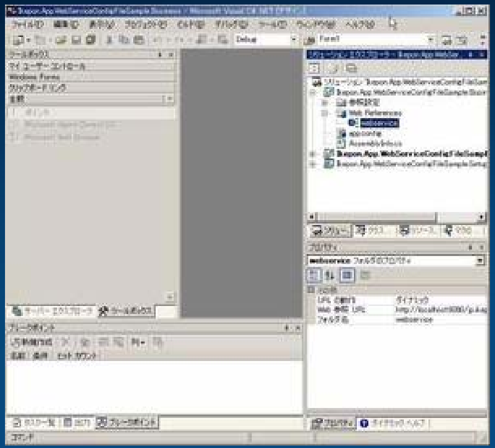
where |Ki–Ai| depends on the type of relation Ki and answer Ai:
- if either Ki or Ai is Unknown then |Ki–Ai|=0.5
- if either Ki or Ai is Not Applicable then |Ki–Ai|=1
-otherwise Ki and Ai are assigned numerical values:
-Yes=1, Sometimes = 2/3, Seldom = 1/3, No = 0

CDV matrix for a single ontology reduced to animal kingdom was initially used to avoid storage size problems.
The first few steps find keywords with IG≈1.
CDV vectors are too sparse, with 5-20, average 8, out of ~5000 keywords.
In later stages IG is small, very few concepts eliminated.

More information is needed in the semantic memory! Active dialogs.



Query

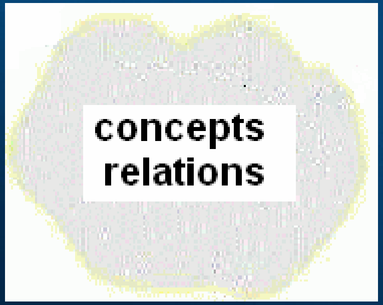


Semantic memory

Applications, eg.
20 questions game

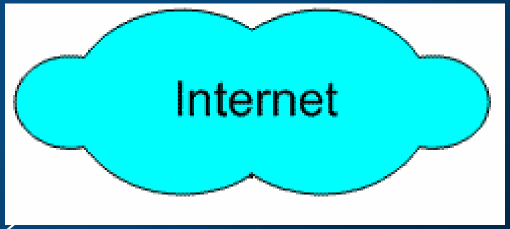
Humanized interface

Store



Part of speech tagger
& phrase extractor

verification

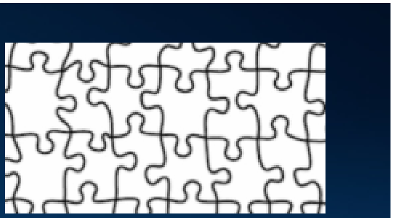


On line dictionaries
Active search and
dialogues with users

Parser

Manual

Puzzle generator



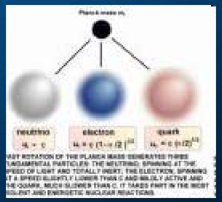
Semantic memory may be used to invent automatically a large number of word puzzles that the avatar presents.
This application selects a random concept from all concepts in the memory and searches for a minimal set of features necessary to uniquely define it; if many subsets are sufficient for unique definition one of them is selected randomly.

It is an Amphibian, it is orange and has black spots.
How do you call this animal?

A Salamander.

It has charm, it has spin, and it has charge.
What is it?

If you do not know, ask Google!
Quark page comes at the top …



Medical applications: goals & questions

- Can we capture expert’s intuition evaluating document’s similarity, finding its category?
- How to include *a priori* knowledge in document categorization – important especially for rare disease.
- Provide unambiguous annotation of all concepts.
- Acronyms/abbreviations expansion and disambiguation.
- How to make inferences from the information in the text, assign values to concepts (true, possible, unlikely, false).
- How to deal with the negative knowledge (not been found, not consistent with ...).
- Automatic creation of medical billing codes from text.
- Semantic search support, better specification of queries.
- Question/answer system.
- Integration of text analysis with molecular medicine.

Provide support for billing, knowledge discovery, dialog systems.

Example of clinical summary discharges

Jane is a 13yo WF who presented with CF bronchopneumonia.

She has noticed increasing cough, greenish sputum production, and fatique since prior to 12/8/03.

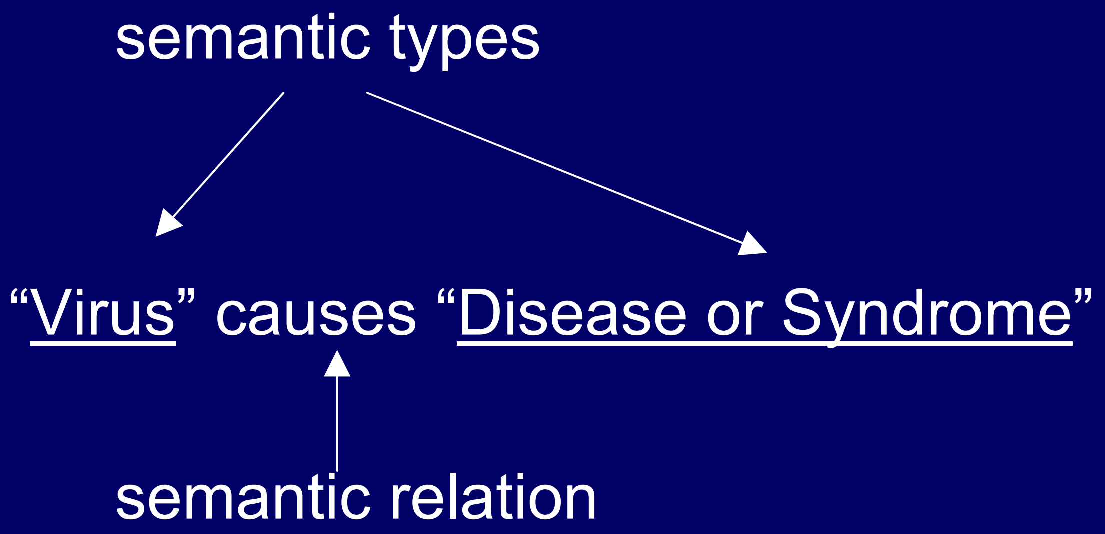
She had 2 febrile epsiodes, but denied any nausea, vomiting, diarrhea, or change in appetite.

Upon admission she had no history of diabetic or liver complications.

Her FEV1 was 73% 12/8 and she was treated with 2 z-paks, and on 12/29 FEV1 was 72% at which time she was started on Cipro.

She noted no clinical improvement and was admitted for a 2 week IV treatment of Tobramycin and Meropenem.

Unified Medical Language System (UMLS)



semantic types

“Virus” causes “Disease or Syndrome”

semantic relation

- Other relations: “interacts with”, “contains”, “consists of” , “result of”, “related to”, …
- Other types: “Body location or region”, “Injury or Poisoning”, “Diagnostic procedure”, …

UMLS – Example (keyword: “virus”)

- Metathesaurus:

Concept: Virus**,** **CUI:** C0042776,
Semantic Type: Virus

Definition (1 of 3)**:**

Group of minute infectious agents characterized by a lack of independent metabolism and by the ability to replicate only within living host cells; have capsid, may have DNA or RNA (not both). (CRISP Thesaurus)

Synonyms: Virus, Vira Viridae

- Semantic Network:

"Virus" causes "Disease or Syndrome"

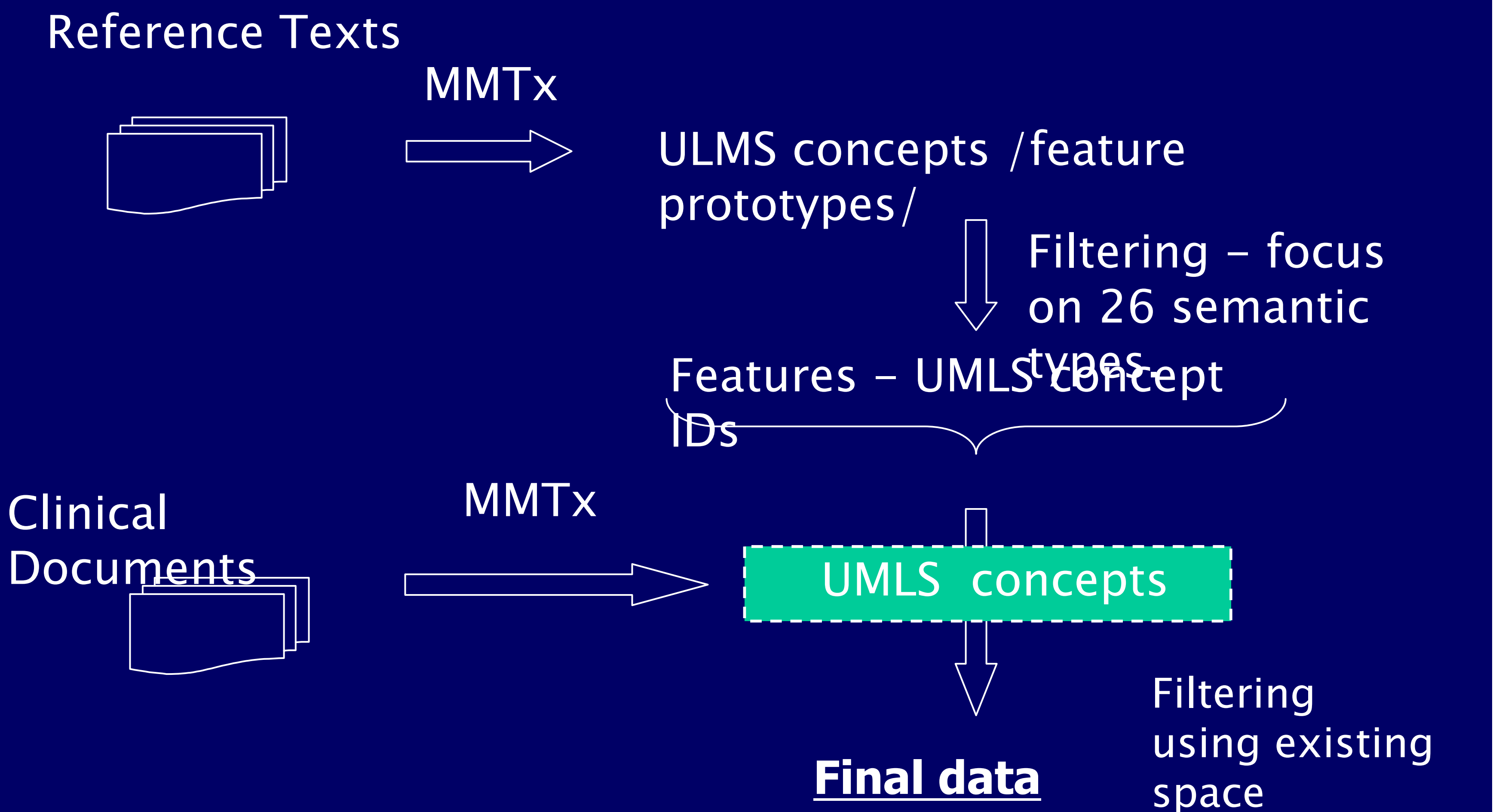
Summary discharge test data

Disease name	Clinical Data		Reference Data size [bytes]
	No. of records	Average size [bytes]	
Pneumonia	609	1451	23583
Asthma	865	1282	36720
Epilepsy	638	1598	19418
Anemia	544	2849	14282
UTI	298	1587	13430
JRA	41	1816	27024
Cystic fibrosis	283	1790	7958
Cerebral palsy	177	1597	35348
Otitis media	493	1420	32416
Gastroenteritis	586	1375	9906

JRA - Juvenile Rheumatoid Arthritis UTI - Urinary tract infection

Data processing/preparation

MMTx – discovers UMLS concepts in text



Reference Texts

MMTx

ULMS concepts /feature prototypes/

Filtering – focus on 26 semantic types

Features – UMLS concept IDs

Clinical Documents

MMTx

UMLS concepts

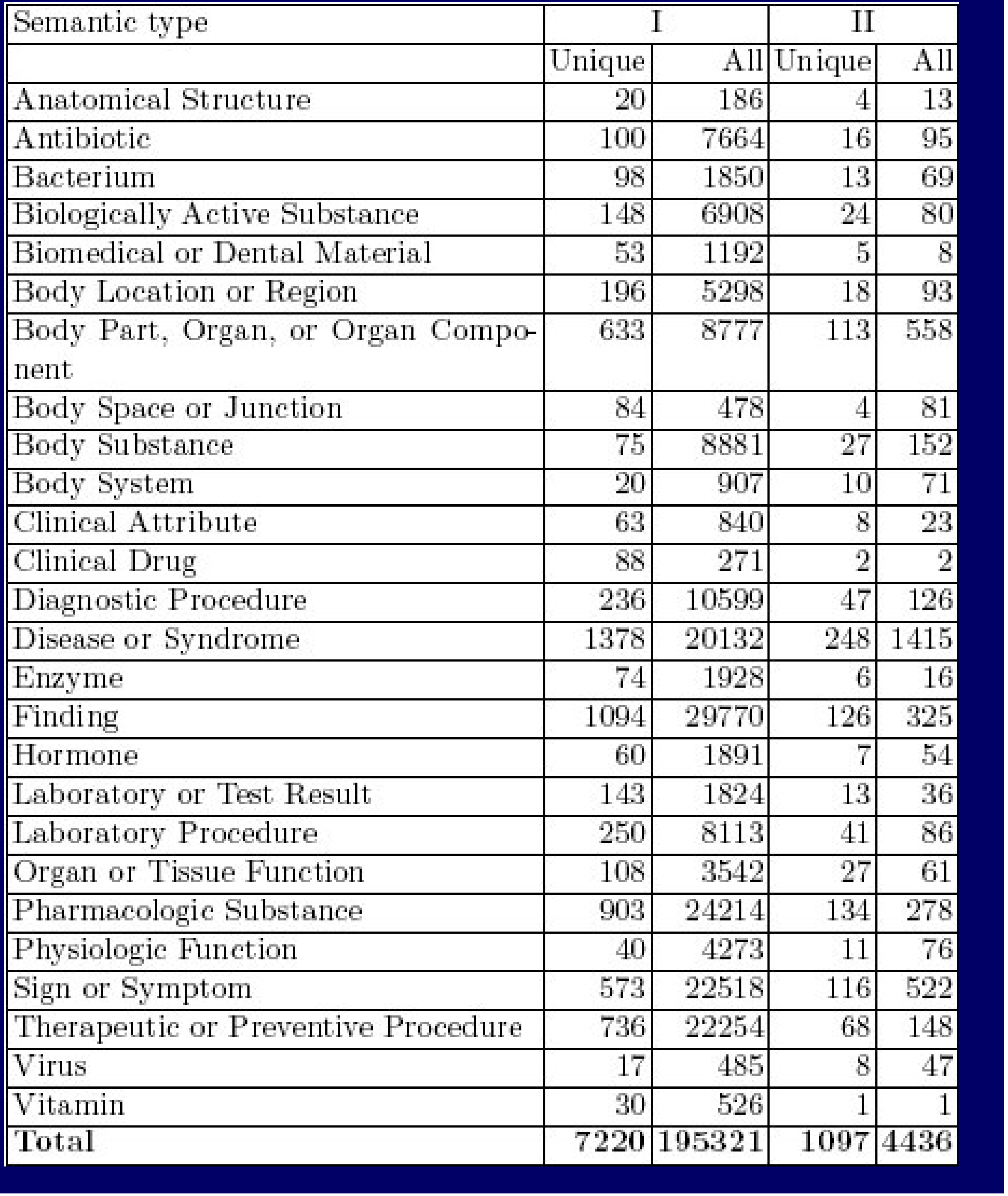
Filtering using existing space

Final data

Semantic types used

Values indicate the actual numbers of concepts found in:

I – clinical texts
II – reference texts



Semantic type	I		II	
	Unique	All	Unique	All
Anatomical Structure	20	186	4	13
Antibiotic	100	7664	16	95
Bacterium	98	1850	13	69
Biologically Active Substance	148	6908	24	80
Biomedical or Dental Material	53	1192	5	8
Body Location or Region	196	5298	18	93
Body Part, Organ, or Organ Component	633	8777	113	558
Body Space or Junction	84	478	4	81
Body Substance	75	8881	27	152
Body System	20	907	10	71
Clinical Attribute	63	840	8	23
Clinical Drug	88	271	2	2
Diagnostic Procedure	236	10599	47	126
Disease or Syndrome	1378	20132	248	1415
Enzyme	74	1928	6	16
Finding	1094	29770	126	325
Hormone	60	1891	7	54
Laboratory or Test Result	143	1824	13	36
Laboratory Procedure	250	8113	41	86
Organ or Tissue Function	108	3542	27	61
Pharmacologic Substance	903	24214	134	278
Physiologic Function	40	4273	11	76
Sign or Symptom	573	22518	116	522
Therapeutic or Preventive Procedure	736	22254	68	148
Virus	17	485	8	47
Vitamin	30	526	1	1
Total	**7220**	**195321**	**1097**	**4436**

Data statistics

General:

- 10 classes
- 4534 vectors
- 807 features (out of 1097 found in reference texts)

Baseline:

- Majority: 19.1% (asthma class)
- Content based: 34.6% (frequency of class name in text)

Remarks:

- Very sparse vectors
- Feature values represent term frequency (tf) i.e. the number of occurrences of a particular concept in text

Model of similarity I

Try to capture some intuitions combining evidence while scanning the text:

1. Initial distance between document D and the reference vectors R_k should be proportional to $d_{0k} = ||D - R_k|| \propto 1/p(C_k) - 1$
2. If a term i appears in R_k with frequency $R_{ik} > 0$ but does not appear in D the distance $d(D,R_k)$ should increase by $\Delta_{ik} = a_1 R_{ik}$
3. If a term i does not appear in R_k but it has non-zero frequency D_i the distance $d(D,R_k)$ should increase by $\Delta_{ik} = a_2 D_i$
4. If a term i appears with frequency $R_{ik} > D_i > 0$ in both vectors the distance $d(D,R_k)$ should decrease by $\Delta_{ik} = -a_3 D_i$
5. If a term i appears with frequency $0 < R_{ik} \le D_i$ in both vectors the distance $d(D,R)$ should decrease by $\Delta = -a R$

Model of Similarity II

Given the document *D*, a reference vector R_k and probability $p(i|C_k)$ probability that the class of *D* is C_i should be proportional to:

$$S(C_k \mid D; R_k) = 1 - \sigma\left(\beta\left[d_{0k} + \sum_i p(i \mid C_k)\Delta_{ik}\right]\right)$$

where Δ_{ik} depends on adaptive parameters $a_1,\ldots,a_4$ which may be specific for each class.
Linear programming technique can be used to estimate a_i by maximizing similarity between documents and reference vectors:

$$d_{0k} + \sum_i p(i \mid C_k)\Delta_{ik} = \min$$

with the constrains:

$$\sum_i p(i \mid C_j)\Delta_{ij} - \sum_i p(i \mid C_k)\Delta_{ik} \geq d_{0k} - d_{0j}; \quad k \neq j = 1\ldots K$$

where *k* indicates the correct class.

Results

	M0	M1	M2	M3	M4	M5
kNN	48.9	50.2	51.0	51.4	49.5	49.5
SSV dec. tree	39.5	40.6	31.0	39.5	39.5	42.3
MLP (300 neur.)	66.0	56.5	60.7	63.2	**72.3**	**71.0**
SVM (Optimal C)	59.3 (1.0)	60.4 (0.1)	60.9 (0.1)	60.5 (0.1)	59.8 (0.01)	60.0 (0.01)
10 Ref. vectors	**71.6**	-	**71.4**	**71.3**	70.7	70.1

10-fold crossvalidation accuracies in % for different feature weightings. M0: *tf* frequencies; M1: binary data;

$$M2: \quad \sqrt{tf}$$

$$M3: \quad 1+\log(tf)$$

$$M4: \quad s_{ij} = 1+\log tf_{ij} \log N / df_i$$

$$M5: \quad s_{ij} = round\left(10 \times \frac{1+\log tf_{ij}}{1+\log l_j} \log \frac{N}{df_i}\right)$$

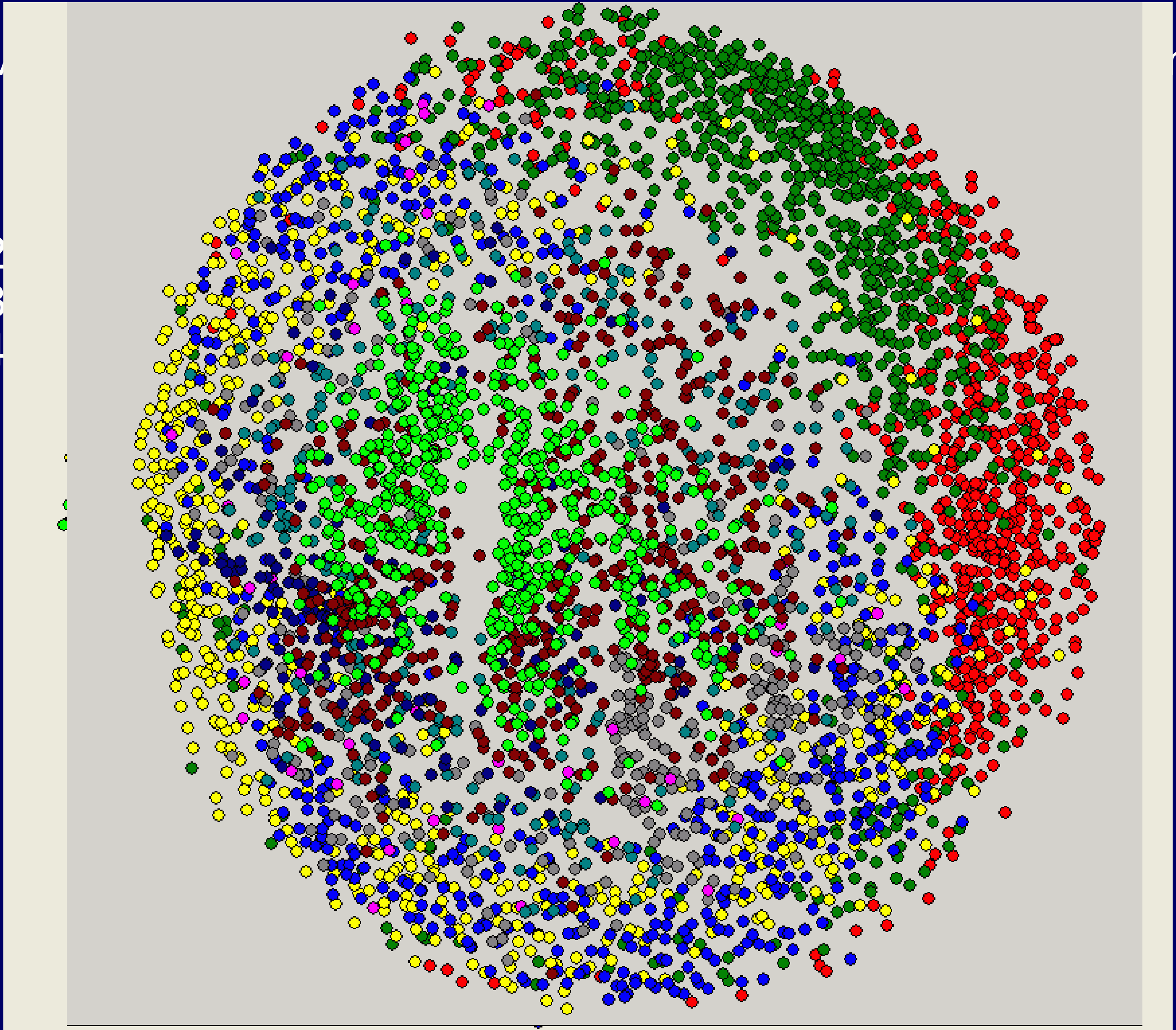
Enhancing representations

A priori knowledge is form of reference prototypes is not sufficient. Experts reading the text activate their semantic memory and add a lot of knowledge that is not explicitly present in the text.

Semantic memory is difficult to create: co-occurrence statistics does not capture structural relations of real objects and features.
Better approximation (not as good as SM): use ontologies adding parent concepts to those discovered in the text.

Ex: IBD => [C0021390] Inflammatory Bowel Diseases =>
-> [C0341268] Disorder of small intestine
-> [C0012242] Digestive System Disorders
-> [C1290888] Inflammatory disorder of digestive tract
-> [C1334233] Intestinal Precancerous Condition
-> [C0851956] Gastrointestinal inflammatory disorders NEC
-> [C1285331] Inflammation of specific body organs
-> [C0021831] Intestinal Diseases
-> [C0178283] [X]Non-infective enteritis and colitis
[C0025677] Methotrexate (Pharmacologic Substance) =>
-> [C0003191] Antirheumatic Agents
-> [C1534649] Analgesic/antipyretic/antirheumatic

Clusterization on enhanced data



M

ng cosine

1

2

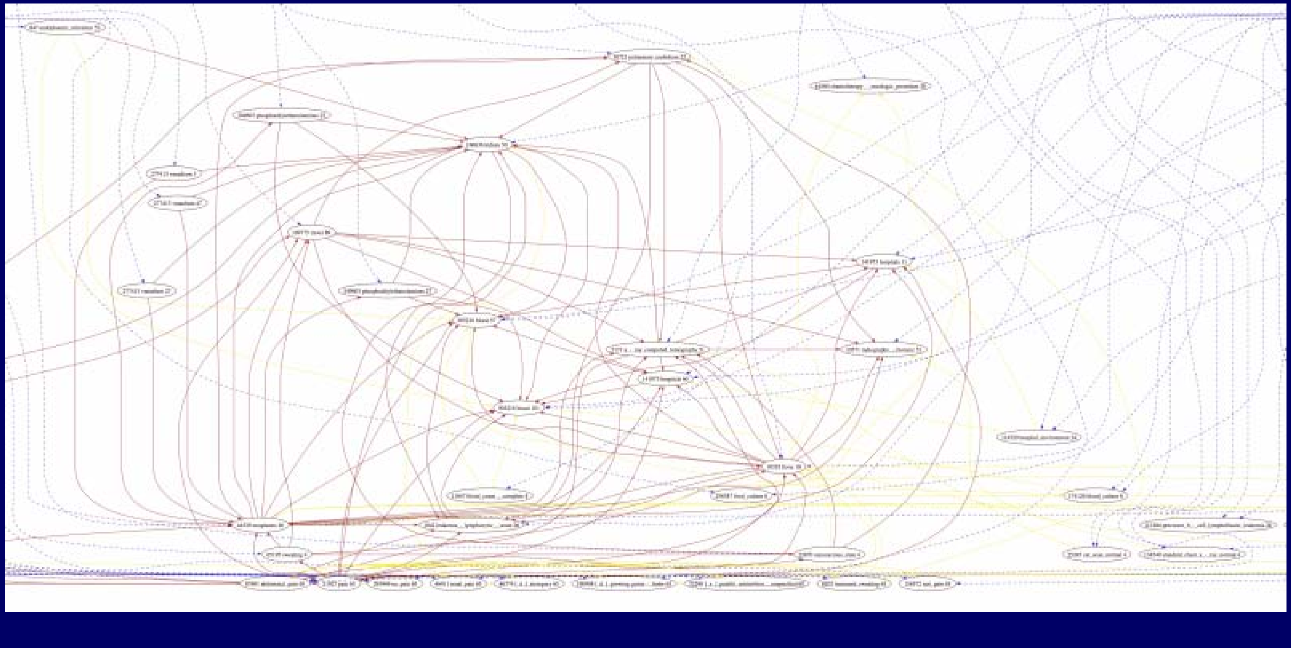
3

4

More semantic relations

Neurocognitive approach to language understanding: use recognition, semantic and episodic memory models, create graphs of consistent concepts for interpretation, use spreading activation and inhibition to simulate effect of semantic priming, annotate and disambiguate text.

For medical texts ULMS has >2M concepts, 15M relations … developing a System for Unambiguous Concept Mapping in Medical Domain (with Matykiewicz, Pestian), and ontology for common reason (with Szymanski)



Graphs of consistent concepts

General idea: when the text is read and analyzed activation of semantic subnetwork is spread; new words automatically assume meanings that increases overall activation, or the consistency of interpretation.

Many variants, all depend on quality of semantic network, some include explicit competition among network nodes.

1. Recognition of concepts associated with a given concept:

1.1 look at collocations, and close co-occurrences, sort using average distance and # occurrences;

1.2 accept if this is a ULMS concept; manually verify if not;

1.3 determine fine semantic types, what states/adjectives can be applied.

2. Create semantic network:

2.1 link all concepts, determine initial connection weights (non-symmetric);

2.2 add states/possible adjectives to each node (yes/no/confirmed

GCC analysis

After recognition of concepts and creation of semantic network:

3. Analyze text, create active subnetwork (episodic working memory) to

   make inferences, disambiguate, and interpret the text.

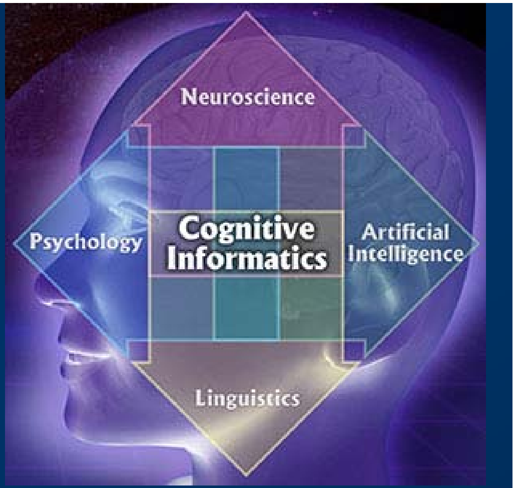
3.1 find main unambiguous concepts, activate and spread their activations within semantic network; all linked concepts become partially active, depending on connection weights.

3.2 Polysemous words, acronyms/abbreviations in expanded form, add to the overall activation; active subnetwork activates appropriate meanings stronger than other meaning, inhibition between competing interpretations decreases alternative meanings.

3.3 Use grammatical parsing and hierarchical semantic types constraints (Optimality Theory) to infer the state of the concepts.

3.4 Leave only nodes with activity above some threshold (activity decay).

Few conclusions



Neurocognitive NLP leads to interesting inspirations (Sydney Lamb, Rice Univ, quite general book).

Creation of novel interesting words is possible at the human competence level, opening a new vista in creativity research and suggesting new experiments.

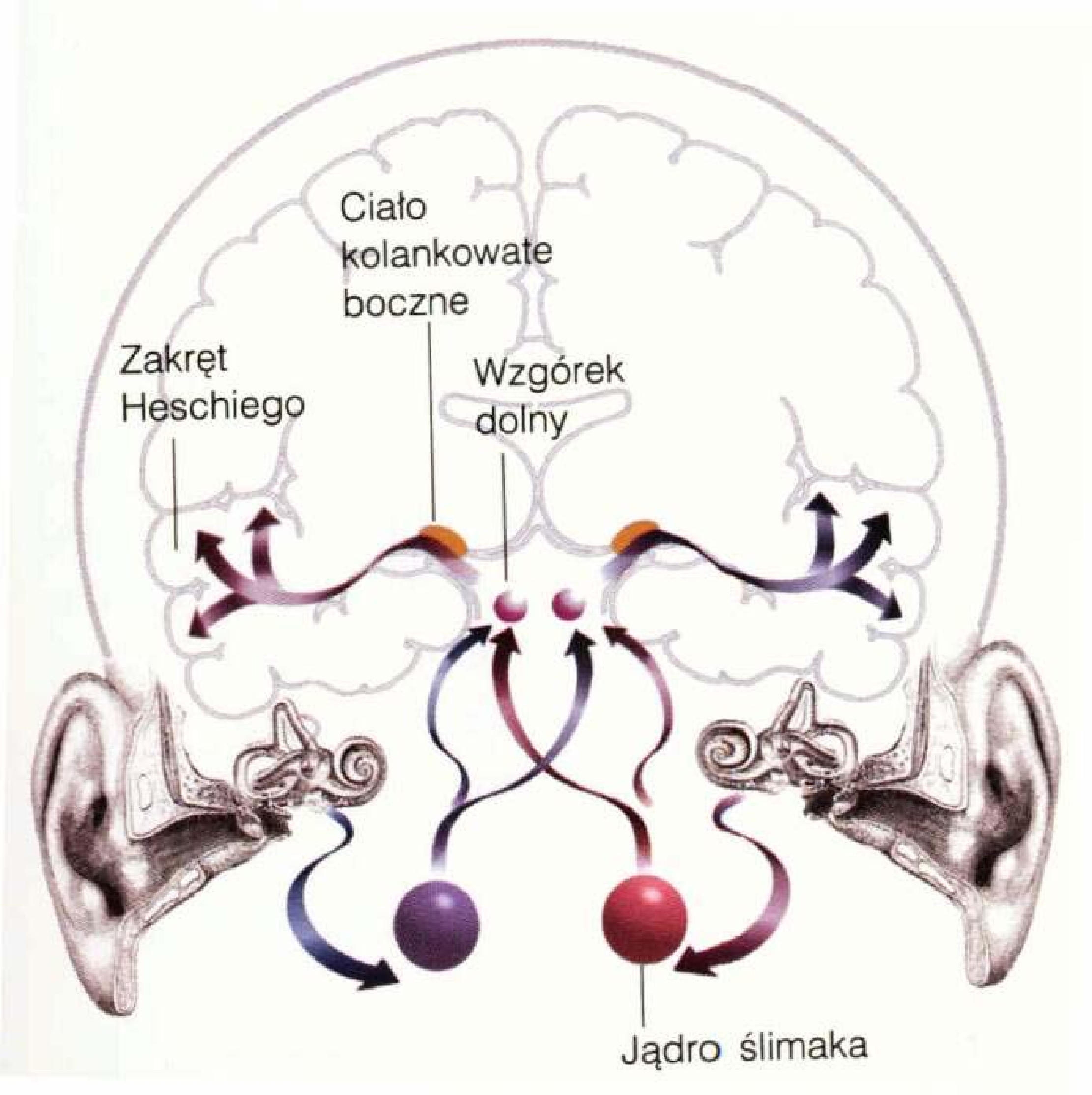
Specific (drastically simplified) representation of semantic knowledge is sufficient in word games and query precisiation applications.

Various approximations to knowledge representation in brain networks should be studied: from the use of *a priori* knowledge based on reference vectors, through ontology-based enhancements, to graphs of consistent concepts in spreading activation networks.

More work on semantic memory for common sense and specialized applications is needed.

Sessions on Medical Text Analysis and billing annotation challenge, April 1-5, 2007, IEEE CIDM, Honolulu, Hilton Hawaiian Village Hotel.

Thank you for lending your ears ...



Google: Duch => Papers