Available online at www.sciencedirect.com
ScienceDirect
Cognitive Systems Research 14 (2012) 84–100
Cognitive Systems RESEARCH
www.elsevier.com/locate/cogsys

Information retrieval with semantic memory model

Action editor: Minho Lee

Julian Szymański [a], Włodzisław Duch [b,c,*]

[a] Department of Computer Systems Architecture, Gdańsk University of Technology, Poland
[b] Department of Informatics, Nicolaus Copernicus University, Toruń, Poland
[c] School of Computer Science, Nanyang Technological University, Singapore

Available online 13 February 2011

Abstract

Psycholinguistic theories of semantic memory form the basis of understanding of natural language concepts. These theories are used here as an inspiration for implementing a computational model of semantic memory in the form of semantic network. Combining this network with a vector-based object-relation-feature value representation of concepts that includes also weights for confidence and support, allows for recognition of concepts by referring to their features, enabling a semantic search algorithm. This algorithm has been used for word games, in particular the 20-question game in which the program tries to guess a concept that a human player thinks about. The game facilitates lexical knowledge validation and acquisition through the interaction with humans via supervised dialog templates. The elementary linguistic competencies of the proposed model have been evaluated assessing how well it can represent the meaning of linguistic concepts. To study properties of information retrieval based on this type of semantic representation in contexts derived from on-going dialogs experiments in limited domains have been performed. Several similarity measures have been used to compare the completeness of knowledge retrieved automatically and corrected through active dialogs to a "golden standard". Comparison of semantic search with human performance has been made in a series of 20-question games. On average results achieved by human players were better than those obtained by semantic search, but not by a wide margin.

1. Introduction

Natural language processing (NLP) techniques that provide effective algorithms for search of relevant information in a huge amount of text documents available in machine readable form are in a growing demand. Search techniques have for a long time been based mainly on keywords. Single keywords or a few keywords (user queries) work well for small repositories of documents that belong to a single domain. More advanced NLP methods are required if search is made in large repositories containing documents from diverse domains. This is due to the strong ambiguity of keywords, leading to low precision, that is returning many unwanted documents, and to the idiosyncratic use of words by different authors, leading to low retrieval rates of relevant information. Effective NLP methods for information retrieval must rely on some basic knowledge about properties of language, and in particular about the semantics of concepts. The knowledge base should approximate relations between lexical elements as a prerequisite to achieve high linguistic competence. The use of such knowledge will be an important step forward towards automation of the process of natural language understanding.

Reading and understanding texts people employ additional background knowledge stored in different types of their memory. Thanks to the recognition memory small mistakes in the texts are ignored, semantic memory associates words with their general meaning in a given context, and episodic memory allows to build model of discourse or narrative. This leads to a rich conceptual view of texts being read that is usually beyond capabilities of NLP systems. A

* Corresponding author at: Department of Informatics, Nicolaus Copernicus University, Toruń, Poland.
E-mail addresses: julian.szymanski@eti.pg.gda.pl (J. Szymański), wduch@is.umk.pl (W. Duch).

1389-0417/$ - see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.cogsys.2011.02.002

lot is know now about brain mechanisms responsible for understanding language (Feldman, 2006; Lamb, 1999; Pulvermnller, 2003). Action-perception circuits in the brain are activated by phonological and visual inputs, and distribution of brain activity provides natural basis for representation of concepts (Duch, Matykiewicz, & Pestian, 2008). These activations change quickly in time and strongly depend on priming by the context (McNamara, 2005) and on previous neural activity, therefore it is not easy to approximate them by knowledge representation schemes used in natural language processing.

The first step to improve NLP methods requires focusing on a better understanding basic concepts represented by words. Understanding here means the ability to give the word its proper meaning in agreement with the context it appears in, and to be able to answer questions that depend on correct interpretation of properties associated with the concept a given word points to. A related aspect of understanding the meaning of the word is the ability to create correct associations relevant in the context of the linguistic episode, giving responses based on information that has not been explicitly given, but may be retrieved from episodic and semantic memory of the cognitive agent. These two basic steps are essential to process natural language in a similar way as humans do.

In the next section psycholinguistic models of semantic memory are briefly reviewed, in the third section our approach to the knowledge representation by semantic memory is presented, Section 4 describes the semantic search algorithm, and Section 5 shows one particular application of this algorithm to the 20-questions game. Section 6 introduces active dialogs for knowledge acquisition, and Section 7 compares results of our algorithm to those achieved by humans playing the same game. The final section contains discussion and presentation of plans for future research.

2. Psycholinguistic models of semantic memory

The idea of semantic memory has been introduced by Tulving, Bower, and Donaldson (1972). He proposed to distinguish memory involved in the cognition process that is used for organization of different types of human experience. Within long term memory structures he distinguished two kinds of memories, called the episodic and the semantic memory. Episodic memory refers to personal experiences, events from the past that may be recalled. Everyone has unique episodic memories, allowing us to understand idiosyncratic references to past events. Although they are formed from similar types of experiences specific configurations of these experiences are always unique. The second is related to the human language system, and is roughly common for all users of a given language, enabling communication process.

Of course any division of brain processes into separate components is only approximate. Both types of memory are simultaneously active. Experiences are stored in episodic memory that engages not only cortical, but also hippocampal structures. Through consolidation process relations and properties of objects are turned into abstract representations, stored in the semantic memory. Semantic memory works as a mental lexicon (Gleason & Ratner, 1997), a dedicated knowledge base storing basic lexical elements – concepts, or "units of knowledge". According to the idea of the Triangle of Reference (Ogden, Richards, Malinowski, & Crookshank, 1949) concepts are used for thinking about the referent. Within the semantic memory structures words serve as labels for concepts that describe elements of generalized experience. Words invoke brain states that encode these elements, enabling communication. Isolated concepts have little meaning – semantic memory contains information about relations between them, so they form conceptual network of elements connected with each other by different kinds of associations. They enable to capture the meaning of words extrapolated from relations to other concepts.

Semantic representation of symbols in the brain has been a matter of extensive research (Pulvermnller, 2003) and thanks to various neuroimaging methods a lot is known about action-perception networks that give an intrinsic meaning to simple concepts. Analysis of fMRI scans shows that for different concepts activation within brain areas devoted to perception, motor manipulation, spatial representation, emotional and self-related regions significantly differs. Despite large individual variance of fMRI signals a prototype brain state of many people may be predicted sufficiently well to distinguish it from about a hundred other concepts (Mitchell et al., 2008). Reading simple stories leads to brain activity that reveals places, characters, subjects and objects of actions, goals, representations for visual exploration and motor activity, simulating in the imagination the events of the story as if they had been perceived (Speer, Reynolds, Swallow, & Zacks, 2009). This in a long run gives a chance to create a natural brain-based basis for representation of concepts in semantic memory (Duch et al., 2008).

A few psycholinguistic models of semantic memory exist. They describe how lexical elements are stored and processed by human brains. Below the main approaches that can be used as an inspiration for building computational model are presented.

2.1. Hierarchical model

Hierarchical model (Collins & Quillian, 1969) presented in Fig. 1. is perhaps the simplest and most natural method to organize concepts. In this approach the predefined is_a relation type organizes concepts (representing natural objects) in the form of a taxonomy tree. Other types of relations (e.g. can_a, has_a) that are useful for building additional associations between nodes, may also exist, but in the hierarchical model they have only informative role. The is_a relation introduces inheritance, with properties of the concepts from higher levels of the taxonomy

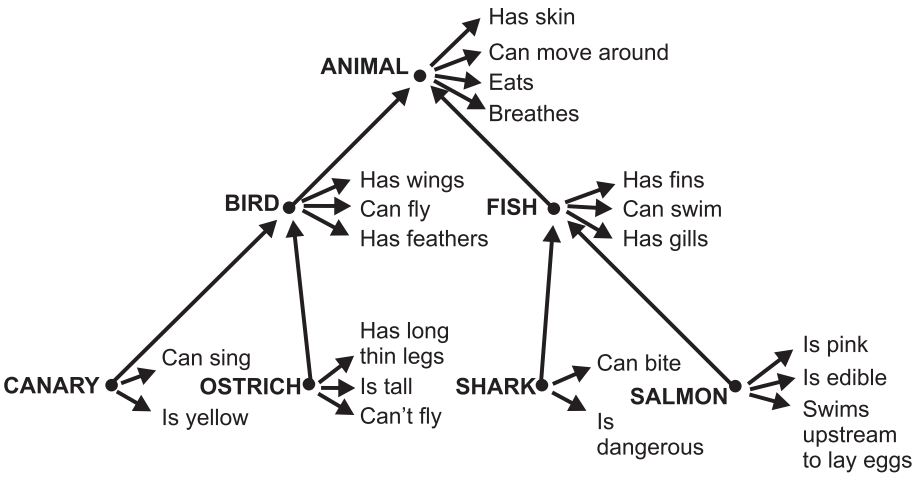

Fig. 1. Hierarchical model of semantic memory.

(parents) being propagated to their children nodes. Most ontologies are based on this type of representation.

The hierarchical model enables to find connections between nodes, and thus provides answers about properties of concepts described by their features. In the simplest case the presence/absence of links is interpreted as the yes or no answer. Studies of response times to simple questions related to properties of objects (e.g. "Is a canary a bird?") revealed that answering questions about typical properties that are directly linked is faster then asking questions about properties that require analysis of links going through the upper taxonomy relations. This is presumably related to the transitions between brain states representing different concepts.

Although ontologies build on hierarchical models have many applications where taxonomy is sufficient, it is clear that the memory structures are not static. Thinking about new relations between two or more concepts that are placed far from each other in the hierarchical tree creates shortcuts or direct associations between these concepts in a way that cannot be accommodated in the hierarchical model. A more realistic approximation to brain processes responsible for acquisition of new semantic knowledge is needed.

2.2. Spreading activation model

The spreading activation model (Collins & Loftus, 1975) of the semantic memory, depicted in Fig. 2, organizes concepts in the form of a lexical network.

Links between nodes of this network describe various relations, including semantic similarities between concepts stored in the network. Concept that is analyzed at a given moment (current thought) is considered to be active, symbolizing coordinated neural activity of many brain areas. If the activation is strong enough it will spread further to several associated concepts triggering their activity. Usually the winner-takes-most neural processes inhibit alternative concepts that could also be activated, leaving only a few that are involved in a sequential thinking process. Spreading activation creates a subnetwork of active concepts associated with the primary concept. In real networks this is a highly non-linear process. Activation of some nodes may result from weak associations with a number of concepts in the analyzed sentence. According to the Hebbian principles frequently used pathways are activated more easily, modifying association strength. Spreading activation to associated concepts depends on the number of hops through intermediate concepts and on their associations strengths, providing a certain distance measure between concepts. Only concepts that are close to the concepts analyzed receive sufficient activation to have some influence on the semantic interpretation. Nodes have finite maximum activations and energy is conserved, therefore nodes with many links may spread only a weak activation, while a few strong associations will lead to larger activations.

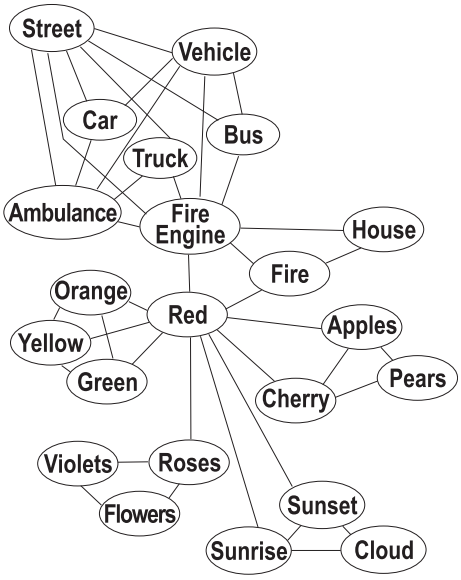

Fig. 2. Spreading activation model of the semantic memory.

This model can describe non-taxonomic similarities between concepts in a better way than hierarchical modeled is able to do. Better approximation to the functions of human semantic memory is seen in tests that analyze concept similarities, where semantic distance between concepts does not increase the time of an answer for false sentences eg.: "all fruits are vegetables" or "fruits are flowers". Empirical studies show that the time of concept activation is related to the semantic distance, measured by intermediate associations that the activation must pass through (Warren, 1977). EEG and fMRI experiments with humans show that associations between closely related concepts arise in 30–100 ms (Mitchell et al., 2008). For distant concepts this time significantly grows (>700 ms), as demonstrated in tests with semantic priming using concepts pairs (McNamara, 2005).

The spreading activation theory may be criticized on several points. First, it lacks different types of relations between concepts, Second, it does not keep cognitive economy, each definition of the concept should be complete, with all important associations defined directly. Third, early models of spreading activation did not included inhibitory associations that suppress concepts associated with a given word, but not suitable in the particular context. Inhibitory associations restrict activation flow to those nodes that represent concepts relevant to the meaning of the whole sentence. This extension of the model has been successfully used for disambiguation of medical concepts in the graph of consistent concepts (GCC) (Matykiewicz, Pestian, Duch, & Johnson, 2006).

Dynamical aspects of biological memory are thus captured to some degree in the spreading activation model, although the processes of forming episodic memories that contribute to the formation of new semantic representations is neglected. A more faithful representation of memory should include also a process of adding and removing

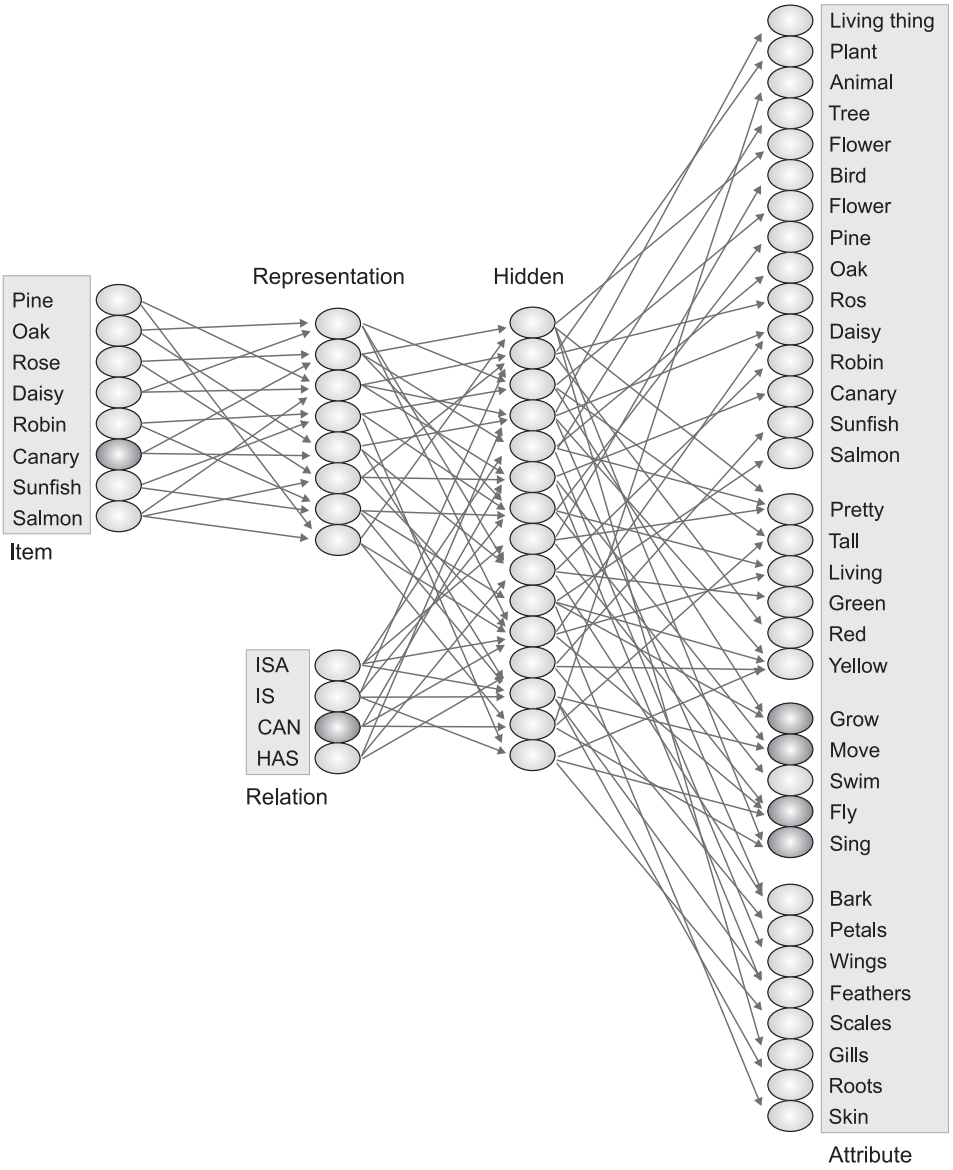

Fig. 3. Semantic memory implemented by a feed-forward neural network (after McClelland and Rogers, 2003).

links as a function of new experiences and forgetting links that have not been activated for a long time.

2.3. Connectionist distributed models

Modern approaches to understanding language processes descend from connectionist models introduced in the parallel distributed processing (PDP) book (Rumelhart & McClelland, 1986). In this approach semantic memory is modeled by a neural network, where the meaning of concepts results from the network dynamics that depends on the connections between neurons involved in distribute representations. In such models information is processed by interactions of many simple elements connected with each other via inhibitory or excitatory links. Distributed representations provided by neural network functions share many characteristics of human memory: they can deal with incomplete or distorted information, display content-based addressing sensitive to context, allow for automatic generalizations, producing similar activation patterns for similar outputs. This allows for application of various types of attributes in retrieval of information stored in memory structures, with features that have strong impact only in precise contexts, for example keywords *bird, does not fly, cold climate* are sufficient to activate the concept *penguin*, while keywords *bird, does not fly, hot climate* the concept *ostrich* or *emu*. In contrast to the connectionist networks information is not localized in a single network node, but is contained in coherent patterns of neuronal activations. This leads to a true sub-symbolic representation of knowledge as single neurons do not represent microfeatures. However, in some simplified neural models identification of a subset of network nodes with microfeatures may be desirable, as it is done in the neural model of human memory shown in Fig. 3.

This model, developed by McClelland and Rogers (2003) for description of natural categories from plant and animal domains, encodes relations of 4 types (ISA, IS, CAN, HAS) between objects and their properties. A feed-forward neural network with two hidden layers learns distributed representations of input objects, using as input plant and animal names and one of the relations, and as outputs properties of these objects. Simple sentences, like "Robin can fly" are parsed to determine inputs, relations and outputs. The final structure stores the knowledge in the form object – relation type – feature. In the learning process layers named *representation* and *hidden* develop internal representations of objects processed by the neural network. This may be seen in the dendrograms showing distributions of hidden layer activity after training. Activity vectors are similar (measured by cosine distance) for each group: trees, flowers, birds or fish, reflecting similarity of objects withing the group, and progressively large differences between plants and animals. Such network shows spontaneous generalizations of information that, although not given explicitly may be induced through similar features of the presented objects. It may also build new associations between categories that have not been given during the learning process. Thus it shows how episodic knowledge, based on collection of facts in form of simple assertions, is converted into semantic knowledge with specific structure.

2.4. Approaches to estimation of meaning

Talking about psycho-linguistic semantic memory models the question how the meaning of concepts is determined by the human brain should be considered. The simplest theories try to explain categorization of natural objects. Thus understanding is simply replaced by assigning a given concept to a proper category, assuming that the meaning of these categories has been established. Two main approaches should be mentioned here.

(A) *Theory of semantic features* (Smith, Shoben, & Rips, 1974).

This theory is based on defining a concept as a list of its features. The features can be divided into two sets:

1. defining features – determining the meaning of the concept,
2. characteristic features – determining the typicality of the concept.

This model takes into account common and differentiating features used during retrieval of similar concepts in the decision process. According to (Smith et al., 1974) conjecture comparison of features is a two stage process. In the first step quick rating of general and typical features is done, allowing for fast decisions. If in this phase similarity of input with the known concepts has not been successfully established slower, more detailed second stage of analyzing defining features is performed. For example, a question "Is canary a bird?" leads to a strong activation based on analysis of characteristic features, allows or quick verification of the truth of this sentence. Verification of the sentence "Is penguin a bird?" is not so direct because features that are characteristic to most birds are missing, therefore a second stage of the defining features comparison should be performed.

Theory of semantic features accounts well for the typicality effects – judgments for typical members of a given category are faster then for unusual members, leading to slower response times (eg. penguin is a bird, vs. canary is a bird). This is explained by the need for a two stage comparison process before the answer is given.

However, empirical studies show some gaps in this theory, called category size effects (Forster, 2004). Analyzing the sentences such as a "poodle is a dog" and a "squirrel is an animal" it has been shown that people evaluate sentences for objects that belong to a narrow (more precise) category faster, despite the fact that precise categories contain more features than higher, more abstract categories

(new specific features are added to the inherited notions). According to the semantic feature theory more comparisons should be performed in this case, taking longer time, contrary to experimental results. Another drawback of the theory is the lack of cognitive economy. It also does not take into consideration the types of relations between objects that may influence the similarity.

(B) *Theory of prototypes.*

Instead of focusing on features this theory of categorization of concepts is focused on the whole objects. Some objects are more typical than others, eg. *chair* is a typical element of the semantic category *furnitures*, more central than for example a *wastebasket*. The prototypes theory (Rosch, 1973) come from psychological research on natural categories, and takes a different approach than traditional thinking in terms of sufficient and necessary conditions. The concept here is not defined by its features but rather by its similarity to a prototype for each category, with unequal category membership status for different objects (e.g. canary is a better prototype of the bird category than penguin). Thus the prototype theory assumes existence of some archetypes representing semantic categories. Objects are assigned to categories using similarity measures performed on different processing levels.

Some evidence of organizing human cognition in the form of prototypes has been given in the research on building artificial conceptual categories (Posner & Keele, 1968). It is clear the prototypes must result from generalization of experience, but it has not been shown how exactly they arise. Similarity functions are usually calculated using a set of feature values, although brains may simply evaluate similarity of distributions of neural activities. A single prototype for each category is not feasible. A set of sufficiently similar examples may be generalized to create prototypes corresponding to similar distributions of neural activity. In the (McClelland & Rogers, 2003) model this may correspond for example to average distributions of activity of hidden layer for general concepts, such as "bird", that the network has not been explicitly trained on. Vectors that describe activity for particular birds will be close to this prototype, with untypical features removing them further from the prototype. Still untypical birds are closer to the prototype for "bird" than for "fish", and both are far from "trees" and "flowers".

These approaches for capturing the meaning of the concepts by the brain are the inspiration for building computational model for processing lexical data. Neural models give certainly very interesting results (Miikkulainen, 1993) but do not scale well. There is still a strong need to create simplified models that capture important properties of neural models but are easier to use from computational point of view. A prerequisite for processing lexical data is the repository for storing lexical knowledge. In the next section knowledge representation used for our implementation of computational model of semantic memory is described, retaining functionalities postulated by psycholinguistic theories.

3. Knowledge representation for semantic memory model

Knowledge representation is one of the basic themes in artificial intelligence. It determines the way how information within machine is stored and processed and what kind of inferences can be performed on it (Davis, Shrobe, & Szolovits, 1993). From the human point of view natural language is the most flexible method for expressing knowledge. It is also the most difficult to formalize in artificial systems. The problem of knowledge representation for natural language is still unsolved and recent trends to connect concepts with action–perception in embodied cognitive systems (Ansorge, Kiefer, Khalid, Grassl, & Knig, 2010) shows how difficult this task may be.

No computer system is able to use language in the way humans do, but there are some implementations that help to improve human – computer interaction. Chatterbots are programs designed to maintain dialog with people. Most of them only mimic linguistic competences without any understanding of the meaning of concepts, therefore they fail to give meaningful answer even to simple questions. Question/answer systems, or information retrieval tasks require more advanced approaches than just template-matching or statistical correlations. Despite a lot of marketing hype behind Wolfram Alpha computational knowledge engine and other such systems answering question is still far from satisfactory.

Flexible method for representation of some aspects of language is based on triples in the form of object – relation type – feature. This method has been employed for modeling data with first order logic (Guarino & Poli, 1995), and has been formalized in popular RDF schemes for ontology implementations (Staab & Studer, 2004). Triples have also been used for building semantic networks (Sowa, 1991) and lexical machine readable dictionaries. Below an extended version of triples will be used for implementation of computational semantic memory model which is in agreement with psycholinguistic observations presented in previous Section 2.

In the standard RDF form learning is possible only by adding or removing triples, making it hard to represent uncertain knowledge. Triples may be considered as links between objects, represented by nodes of semantic networks, and features of these objects, with relation determining the type of the link. The simplest way to extend flexibility of triples and enable learning during knowledge acquisition process is to add weights estimating strength of relations. Such weights should encode fuzzy knowledge, to which degree some features are present (conveniently expressed in terms of fuzzy sets Zadeh, 1996), as well as handle uncertainty of knowledge, estimating reliability or typicality of features.

In Fig. 4 the elementary atom of knowledge in the vwORF representation used for implementation of

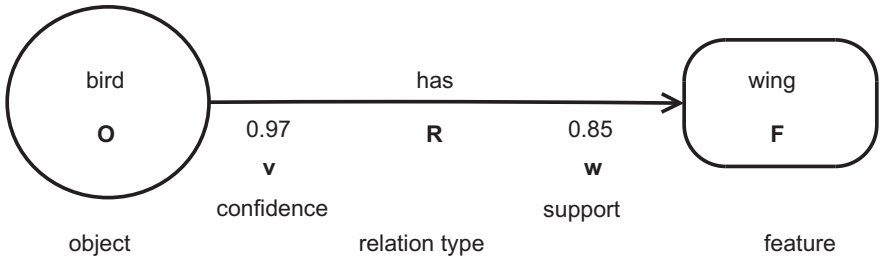

Fig. 4. Atom of knowledge vwORF used in implementation of our semantic memory model.

Semantic Memory computational model is presented. This atom of knowledge consists of five elements which can be divided in two groups:

Triple of knowledge:

O – the object described, represented by its name, usually a concept name rather than a single word.
R – relation type denotes in what way the object is related to the feature.
F – feature or a given property of the object.
Weights:
v – confidence, a real number in the range $\langle 0,1\rangle$, describes how reliable is the knowledge described by this triple. The value of v grows to 1 when knowledge expressed by the triple is repeatedly confirmed; for new triples when there is no confirmations this value may be set near 0.
w – support, a real number in the range $\langle -1,+1\rangle$, describing how typical is this feature for this object. Using this parameter adjectives such as: "always", "frequent", "seldom", "never" can be expressed, for example feature *black* as the property of the *stork* has $w=-0.5$, meaning that it is *seldom* true, and feature *white* should have $w=0.9$, which means that stork is *almost always* white.

With the use of vwORF knowledge representation the meaning of elementary natural language sentences can be expressed. In Fig. 4 an example sentence "bird has wing" has been expressed using vwORF notation. Confidence of this triple ($v=0.97$) is very high, which means that this knowledge has been confirmed through many observations (lifetime of the system). The support ($w=0.87$) is also high, expressing the knowledge that birds generally have wings.

Note also the limitations of such representation: there is no opportunity here to expresses numerical information, for example that a bird has no more than two wings. However, this knowledge can be added by another triple, connecting "bird" and "pair of wings". Some knowledge is expressed more easily by setting constraints rather then specifying precise values. Confidence and support could be functions of numerical values, estimating how likely is a given value (for example hight) for a given object (for example human). We shall not discuss such extensions here.

The set of weighted triples allows to express relatively wide knowledge, including negative knowledge. The set of triples joined together forms semantic network, denoted here with the ζ symbol that represents the whole knowledge stored within the semantic memory model. This knowledge may be represented in a graphical form by visualization of the semantic network. A user friendly interface for navigation over such data using interactive components has been implemented by us, allowing to traverse the graph of concepts and features. This method of visualization has been also used in our other projects[1]: for building WordNet in a cooperative way (Szymański, Dusza, & Byczkowski, 2007) and integrating it with Wikipedia[2].

Presenting knowledge ζ in the form of semantic network is convenient for people, facilitating easy modification using visual interface. Unfortunately data stored in this way cannot be efficiently processed by machines. To enable fast numerical operations semantic network is replaced by geometrical representation called "semantic space" and symbolized by ψ. Turning knowledge ζ contained in semantic network into representation of the semantic space ψ transforms each object node **C** into n-dimensional feature space **F**, where each object is represented by a point, equivalent to a sparse vector of feature values. Many vwORF nodes are defined for each object and they are collected together in the vector called **C**oncept **D**escription **V**ector (CDV).

Exact mapping used here requires two dimensions for each feature to store v and w weights. In its simplest form CDV vectors could store only binary information about existing relations; an intermediate solution is to keep a single real feature value. The number of all features in ζ is large and the number of features that are applicable to a given object is rather small, therefore the vectors are quite sparse. Although some information is lost in such transformation from ζ to ψ it is possible to perform some inferences on knowledge stored in semantic network, thus expanding knowledge that is stored in explicit way in the semantic space. Inferences are based on the processing of the predefined relation types and they add additional features stored in CDV. Four types of relations that appear between Semantic Network nodes are processed:

[1] http://wordventure.eti.pg.gda.pl.
[2] http://swn.eti.pg.gda.pl.

1. **is_a** – The relation introduces in ζ a hierarchy of concepts that facilitates cognitive economy by inheritance of features. If relation of the O_1 **is_a** O_2 type between two objects has been identified features from the CDV(O_2) are copied to the CDV(O_1). A single weight is stored, obtained by multiplication of the v confidence value for the *is_a* relation by the w support value for each feature copied. For all types of relations features that already exists in the CDV of the object are not changed.
2. **similar** – If O_2 is similar to O_1 features from CDV(O_1) should be copied to CDV(O_2), adding additional features that have not been present in CDV(O_2), with the weight factor obtained by multiplication of confidence v related to the relation *similar*, multiplied by the w support for the O_1 features. This relation is not symmetric. However, if $v = 1$ relation *similar* becomes *same*, implementing equivalence of semantic memory objects, therefore processing is performed also from O_2 to O_1.
3. **excludes** – Processing of this relation is similar to the one presented above, except that the w support value of the feature copied to CDV(O_2) is multiplied by -1.
4. **entail** – If F_1 entails F_2 feature F_2 may be added to the CDV of the object for which F_1 is defined, with the w value of the F_2 being the same as F_1, and confidence factor v associated with the relation.

As an example consider semantic network constructed for 172 animals (or more formally, objects from the animal kingdom domain). The 475 features describing them were selected from relations of these objects that have been found in three lexical databases: WordNet (Miller, Beckitch, Fellbaum, Gross, & Miller, 1993), MediaMIT ConceptNet (Liu & Singh, 2004), Microsoft MindNet (Vanderwende, Kacmarcik, Suzuki, & Menezes, 2005). Usage of three different data sources allows to build lexical semantic network in the automatic way. To assure the quality of the knowledge v values have been set using confirmation of particular atoms of information in different sources. Only those relations that appear in more than one data source have been imported, with confidence factor $v = 0.5$ if they are only in two sources, and 0.75 if they are in all three sources. The confidence value associated with each relations is further increased or decreased as a result of the interaction with human user. Knowledge acquired by aggregating three machine readable dictionaries consisted of 5031 most reliable relations describing 172 animals with 475 features.

Performing inferences based on these 4 types of relations enhances CDV representation of objects with new features. Fig. 5 presents how processing a particular relation type during ζ to ψ transformation influences the average number of the features in the CDV vectors.

4. Semantic search algorithm

Semantic network describing relations between lexical elements can be useful in many applications. We have successfully applied the knowledge about relations of the natural language elements, encoded using knowledge representation proposed in previous Section 3, for improvement of text classification (Majewski & Szymański, 2008). Semantic space allows to perform semantic searches for objects of interest referring only to their features. This kind

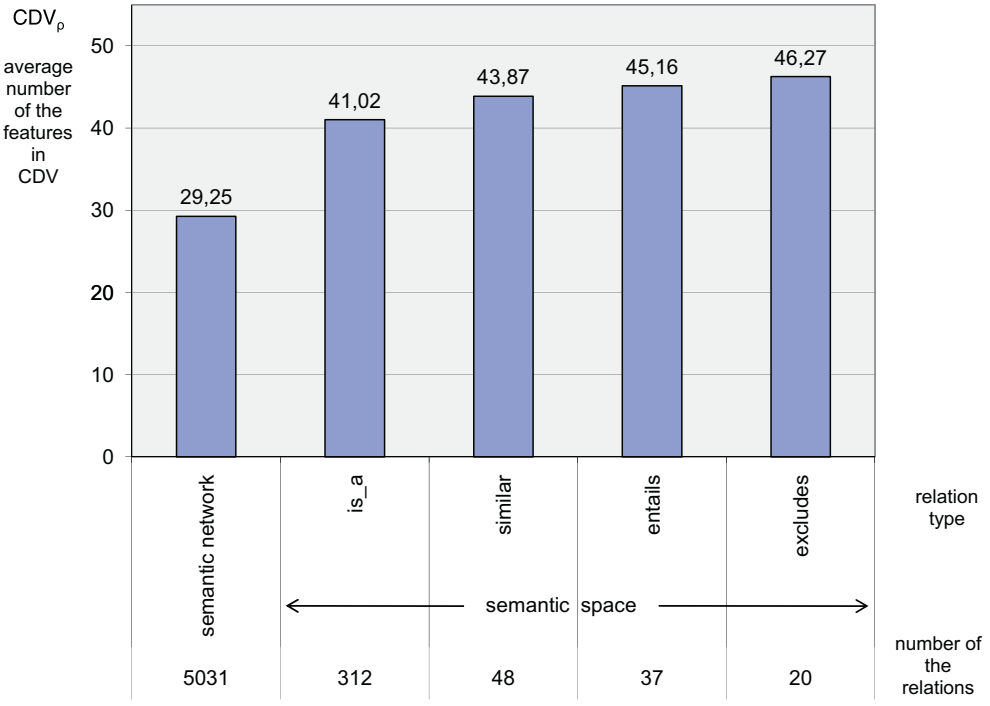

Fig. 5. Change of the average number of specified CDV features as a function of the processed relation types.

of search is useful when a user cannot recall the name of an object she/he is looking for (consider the Tip of the Tongue problem Burke, MacKay, Worthley, & Wade, 1991) or even does not know its proper name. This situation is quite common, for example seeing an image of a flower one may try to identify its name. In such a case typical keyword-based search may not be effective and could be replaced by semantic memory guided search, as described below.

Searching for unknown object in the semantic space is performed by selecting the most distinctive feature that will divide the whole space in a balanced way. In the semantic space ψ containing M objects O_i with N features F_k the search based on a set of keywords should select the best feature that gives the maximum amount of information. Information-based measures are frequently used in decision trees (Quinlan, 1986) and information selection. In our case calculation of information associated with each feature in ψ is done according to the modified Shannon formula (1).

$$I(F_k) = -\sum_{i=1}^{M} \frac{|w_{ik}|}{M} \log \frac{|w_{ik}|}{M} \tag{1}$$

where w_{ik} is the support assigned to the relation between the object O_i and feature F_k. Information (1) depends on the objects contained in the part of ψ space considered, and this subspace is reduced after each new keyword value is specified. For large semantic space ψ it is quite likely that there will be more than one feature having the same highest information, therefore additional heuristic criteria are needed for ranking. One heuristics is to select the most prevalent feature, according to the term frequency stored in databases derived from general corpora (Hunston, 2001). Another heuristics is to use probabilities $p(O_i)$ from previous searches to guess which objects are most frequently searched for. For statistics based on N_S previous searches minimum probability for rare objects selected once is $p(O) = 1/N_S$. Ordering m objects selected out of all M objects in decreasing sequence of $p(O_i)$ probabilities one obtains a curve that has roughly exponential shape. This curve can be used to estimate probability of the remaining $M - m$ objects that have not been selected so far. Calling this probability p_r one may then renormalize estimated probabilities $p(O_i) \leftarrow p(O_i)/(1 + p_r)$ and use them in the modified formula (1):

$$IW(F_k) = -\sum_{i=1}^{M} p(O_i)|w_{ik}| \log p(O_i)|w_{ik}| \tag{2}$$

The best separating feature selected on the basis of $IW(F_k)$ value is used as a keyword. The user determining its value can narrow the set of the objects which can be the result of her/his search. In the implementation presented below we allow only "yes", "no", "don't know", "sometimes" and "frequently" answers, but depending on the application other answers could be accepted (for example, a value in a given range). All answers given by the user are collected in the vector A that is used to calculate distance to all objects in the semantic space and select the most probable (closest) objects. Full representation of object features stored in the $V_i = CDV(O_i)$ is used to calculate Euclidean distance in a subspace of K known feature values:

$$D(A, O_i) = \sqrt{\sum_{k=1}^{K} (V_{ik} - A_k)^2} \tag{3}$$

where K is the length of the answer vector A, describing how many features have been tested so far. CDV vectors do not have full information about relations between objects and features, some answers may not be correct and distances may be influenced by different number of features defined in each CDV vector. Therefore instead of Euclidean distance it is better to use cosine measure, a normalized dot product of the A and V vectors, which has proved to be quite reliable in information retrieval problems (Qian, Sural, Gu, & Pramanik, 2004).

$$d(V, A) = \frac{\sum_i V_i A_i}{\sqrt{\sum_i V_i^2}\sqrt{\sum_i A_i^2}} \tag{4}$$

In our system knowledge has different confidence factors (v), and has fuzzy support (w), therefore instead of these simple measures similarity is computed by:

$$S(V, A) = \frac{1}{K} \sum_{i=1}^{K} (1 - \mathrm{dist}(V_i, A_i)) \tag{5}$$

where the distance between components is defined by:

$$dist(V_i, A_i) = \begin{cases} 0 & \text{if } w(A_i) = \mathrm{NULL} \\ -|w(A_i)|/K & \text{if } v(V_i) = 0 \\ v(V_i)|w(V_i) - w(A_i)| & \text{if } v(V_i) > 0 \end{cases}$$

where $v(V_i)$ and $w(V_i)$ are confidence and support weights describing relations with feature F_i in CDV, and $w(A_i)$ is the answer given by the user to the question "is the feature F_i true" for the object she/he is searching for. The following table shows numerical values that corresponds to the verbal answers:

Similarity of the CDV and answer vectors A is calculated as a sum of differences between user's answers and the system knowledge. If the user answers "don't know" this feature is omitted during calculation of similarity. Additionally the confidence factor v allows to strengthen importance of CDV components which are more reliable and weaken the influence of the accidental ones.

After k steps (answers) maximum similarity $S_{\max}$ between the current answer vector A and all CDV vectors V is achieved in a subspace $\mathcal{O}(A)_k$ containing objects that – with high probability – are looked for:

$$\mathcal{O}(A)_k = \{O_i | S(A, V(O_i P)) = S_{\max}\} \tag{6}$$

Using the maximum similarity or equivalently a minimum distance criteria to construct $\mathcal{O}(A)$ subspace should lead to fastest recognition of the searched objects using

minimum number of questions asked during the search. However, knowledge stored in CDV vectors is not perfect and answers given by the user are not always correct, therefore such approach would sometimes miss searched objects. Moreover, it will miss the opportunity to acquire new knowledge, as discussed further below.

5. The game of questions

Given a limited description of the object people are able to identify it because their semantic memory, storing many relations between objects in the real world, is able to formulate good questions and make inferences that complete partial descriptions. The anticipation and associations created by the lexical context are one of the most important processes that the system capable of understanding natural language in a similar way to humans should posses. It requires good models of semantic and episodic memories.

Semantic search process is a good model of the popular 20-question word game, where one person is allowed to ask 20 questions trying to guess the concept the opponent has in mind. The game is relatively simple for the people, because they have wide knowledge about the world. It is quite difficult for computer programs because success does not depend that much on computational power (as in chess and other board games) but relies on knowledge about the world represented as relations between lexical elements. For that reason it has been proposed as a good challenge for computational intelligence (Duch, 2007). This game may also serve as a demonstration of the elementary linguistic competences based on lexical knowledge, allowing for a real understanding of the meaning of discourse, and not just to respond mechanically using templates.

In our implementation of the game the computer, using knowledge encoded in the form of semantic network, tries to guess the object a human user has in mind. In reply to the given questions, generated from vwORF knowledge representation, a human can give answers only in the form specified in the Table 1. What makes this approach[3] different compared to the ones already available in the Internet[4,5,6] is the flexibility of the knowledge representation used. All knowledge is stored in the semantic network and converted to the vector-based semantic space to increase computational efficiency. This makes our approach largely independent of particular applications. Various applications may use knowledge contained in the semantic network. For example, automatic generation of riddles for crosswords is easily achieved by selecting small subsets of features that allow for a unique identification of objects. A very large number of such subsets exist even in knowledge bases of modest size.

[3] http://diodor.eti.pg.gda.pl.
[4] http://www.20q.net.
[5] http://www.braingle.com/games/animal/index.php.
[6] http://en.akinator.com/.

Ability to use linguistic data in many applications allows to place our approach in the field of artificial general intelligence (Voss, 2005). Alternative approaches to word games encode the knowledge in a fixed form, using a matrix of objects and questions, which makes it easier to process by computers, but is only a superficial imitation of natural language abilities, although still better than the famous ELIZA template-based approach (Weizenbaum, 1966). Another difference is the way in which questions are generated during the game. Other approaches used hand-coded questions, while in semantic search questions are automatically generated using atoms of knowledge in vwORF representation. Formulating questions that are grammatically correct is a challenge in itself because there are many forms (depending on the relation type) in which question could be cast.

The third difference between semantic search approach and other systems is the way knowledge is acquired. This is the main bottleneck of most knowledge-based systems (Cullen & Bryman, 1988). Our implementation bootstraps itself on knowledge from available machine readable dictionaries and other electronic sources, and thus can be run on a large scale, while other projects mostly exploit the interaction with the users to learn correct answers. Of course human–computer interaction is a very useful way of acquiring knowledge, but it is also very time consuming and needs "the snowball effect" to bring enough players, which requires strong marketing. Focusing on automatic data acquisition interaction with humans is used here only for validation and correction of the results, as discussed below in the Section 6.

To make the game of questions more attractive some modifications to the semantic search algorithm are introduced.

1. Questions are generated selecting vwORF atoms from the semantic space according to the formula (1) or (2). If the same user repeats the game searching for the same object several times the deterministic system would ask the same questions, and that could be annoying. This situation is a good opportunity from the knowledge acquisition point of view (see Section 6). To prevent choosing many times the same question stochastic elements are introduced, selecting features randomly with probability related to the information calculated according to (1) or (2). This method is analogous to the popular genetic algorithms roulette reproduction approach (Goldberg, 1989). Such modification makes selecting features (questions) a bit less effective, but in the tests it has not shown significant negative influence on the average number of questions.
2. In the classic version of the semantic search algorithm subspace $O(A)$ containing most probable objects that are used to estimate information has maximum similarity or a minimum distance between the current answer vector and all vectors in $O(A)$ (Eq. (5)). If there is some significant discrepancy between these answers and

stored knowledge (due to the errors in answers, semantic network, or both) the object of interest may be left outside of $O(A)$. To prevent this situation larger subspace is taken, with objects accepted with probability given by modified Boltzman distribution:

$$p(\Delta d, k) = \frac{2}{1 + \exp\left(\frac{k\Delta d}{c}\right)} \tag{7}$$

where k is the step of the game (number of question asked so far), $\Delta d = \mathrm{dist}(V(O), A) \geqslant 0$ is the distance of the CDV vector representing some object O from the answer vector A minus d_{min}, and c is a scaling constant, set to five in all experiments. All objects at minimum distance are always included ($p(0,k) = 1$), while objects that are further then $d_{\min}$ are included with decreasing probability, and may be different then those selected in the previous step. Selecting objects to the subspace $O(A)$ with a little less restrictive criteria than d_{min} makes the game longer, but it can only be observed for popular objects that can be found asking a few questions. For longer games k in the Eq. (7) grows, and only objects at a minimal distance have a chance to be selected. If all answers are correct and match knowledge in the semantic space $d_{min} = 0$, but the subspace $O(A)$ may still contain many objects and more questions will be generated.

3. Algorithm stop condition: three cases when that algorithm may stop are considered:
   - The algorithm stops when there is only one object left in the subspace $O(A)$. It is the most desired situation, but it is relatively seldom because the CDV vectors are sparse – the knowledge relating features with objects is usually far from complete. Also expanding the subspace $O(A)$ using Eq. 7 brings into consideration less probable objects.
   - The algorithm stops after asking the maximum number of questions allowed. Assuming only binary answers to the questions, and minimal differences between objects (objects differ only by one feature), 20 steps (questions) should distinguish $2^{20} = 1{,}048{,}579$ objects. These assumptions are of course not true, in practice knowledge in CDV is incomplete, vectors differ in more than one feature, features are not binary, and more than 20 features are used to describe objects. Notwithstanding these issues, 20 questions seems to be a reasonable maximum number of questions for one game.
   - When the number of objects left in the $O(A)$ subspace is relatively small heuristics may be used to identify searched object. The implemented heuristic is based on the observation that if there exists an object which significantly differs from other objects in the $O(A)$ subspace, and it stands out during successive questions that it may be the object of interest. The implementation of this heuristic is based on fulfilling the condition described by (8).

$$d_p = \Delta(d_{min+1} - d_{min}) > \mathrm{std}(O(A)) \tag{8}$$

Fig. 6. Avatar used in the implementation of the game of questions seen in the Internet Explorer.

where $d_{\min}$ is the minimal distance in the $O(A)$ between CDV and the answer vector, $d_{\min+1}$ is the second minimal distance, $\mathrm{std}(O(A))$ is the standard deviation of distances in the $O(A)$ subspace.

Tests of this heuristic show that it considerably decreases the number of the questions, but in some cases leads to a wrong guess. The tradeoff between the number of questions and precision of finding the object is analogous to that between precision and retrieval – the two measures behave in opposition to each other and there is a problem of optimizing them simultaneously (Buckland & Gey, 1994).

Technical implementation of the game has been done in form of a server controlling a web page with interactive user interface in the form of HIT (**H**umanised **I**n**T**erface). Due to the MS ActiveX technology used for the Avatar, full interaction is possible only under Internet Explorer. Interface in the form of a humanoid taking head is depicted in Fig. 6. This implementation serves as the testbed for integration of technologies making the web applications more user-friendly (Szymański, Sarnatowicz, & Duch, 2007). The Haptek[7] 3D head has been integrated with the text to speech engine (TTS) and endowed with the speech recognition[8] (due to the unacceptably high consumption of server's computational resources available only in the console version). The problems faced with implementation of these attractive technologies on a large scale shows that although the HIT functionalities are implementable they are still not mature enough to be widespread.

[7] http://www.haptek.com.

[8] MS SpeechAPI http://www.microsoft.com/speech/speech2007/default.mspx.

6. Knowledge acquisition through active dialogs

It cannot be expected that a semantic network ζ built automatically from available data sources will be complete, and each object will be properly related to all features that describe it. A method for validating and correction the data acquired automatically is needed. The approach described here is based on the semantic search algorithm implemented in the form of a word game performed by the human user, who modifies the lexical knowledge base as a result of her/his search. The interaction with the program has been limited so far to the answers (in the form defined in the Table 1) accepted to the questions generated using the knowledge stored in the system. Human–computer interaction during the games is enriched through active dialogs based on the templates of interactions, run in a specified parts of the game. Three such templates are described below:

1. At the end of the game, if the system correctly guessed the concept, additional question *Is that right?* has been added to verify quality of knowledge stored in the semantic space. Using the *yes/no* answer given by the user to that question precision of the search is defined as $Q = N_s s/N$, where $N_s s$ denotes the number of the searches that finished with success, and N denotes the total number of the performed searches.
   For the initial semantic network constructed in an automatic way $N = 30$ test searches have been performed for an object randomly selected from ζ set. Selection of object for searches has been done with probability distribution given by a normalized number of the features in CDV vectors, so objects that are better described and more popular are favored. The $Q = 0.7$ result indicates that in the limited domain there are some possibilities to obtain common sense knowledge in the form of relations between lexical concepts automatically. It also shows that the method of integrating semantic data from three machine readable dictionaries requires manual validation and correction.The user's answers given to the questions asked by the system allow for correcting and also obtaining new knowledge stored in the semantic network. The answer vector is used to perform modifications of the knowledge according to the results of the search:
2. If to the last question *Is that right?* a user gives an answer *yes*, the entries in the answer vector are used for enriching the CDV representation of the object the system guessed correctly. If some features present in the answer vector already exist in the CDV the w weights are modified taking the average value of w associated with particular feature in CDV and in the answer vector. In addition the system asks an open question: *Tell me something about* ⟨*found object*⟩. The answer may link existing feature to the object through some type of relations, but also may add a completely new feature to the semantic network that has not existed in the knowledge base. An automatic search for possible links between new feature and stored objects is performed.
   This procedure requires deep linguistic parser to convert the sentence in natural language given by the user to knowledge vwORF knowledge representation (Szymański et al., 2007). Parsing sentences given by the users is the opposite process to the generation of questions performed by the system during the game.
3. If instead of the confirmation of search results the user answer is *no* the system asks additional question: *Well, I fail to guess your concept. What was it?*. The name of the object the user was thinking about indicates which CDV should be corrected according to the information in the answer vector. If the object has not existed in the semantic memory before a new object is created with initial features copied from the answer vector. This active dialog allows the system to learn new objects.

Table 1
User answers and their numerical encodings.

1	For the answer "yes"
0.5	For the answer "frequently"
−0.5	For the answer "seldom"
−1	For the answer "no"
0	Denotes answer "don't know"

To validate active dialog approach five test objects with the largest number of features defined in CDV have been selected. After their manual verification (using interactive visualization of the semantic network) their CDVs were take as the Golden Standard and used to verify capability for acquiring new knowledge through the active dialogs. The verification has been performed by removing these objects from the knowledge base and then learning about these objects through the interaction with users. The averaged dynamic of this process, performed for five objects in ten games has been presented in Fig. 7. All games have been limited to twenty questions.

The process of acquiring knowledge using active dialogs has been monitored analyzing how complete the CDV of a new object (NO) became comparing to the Golden Standard (GS). To analyze the process of acquiring knowledge four measures are introduced:

1. $S_d = N_f(GS) - N_f(O)$ is the measure of incompleteness of the new object, showing how the NO differs from GS in terms of the number of features. $N_f(GS)$ is the number of features defined in the Golden Standard $GS = G(O)$ for the concept O, and $N_f(O)$ is the number of features defined for this concept in $CDV(O)$. The S_d value shows how many features are still missing compared to the golden standard.
2. $S_{GS} = \sum_{i=1}^{N_f(O)} [1 - \delta(CDV_i(GS) - CDV_i(O))]$ is the measure of similarity based on the co-occurrence of features. It shows more precisely than S_d how close is NO to GS. The sum is only over features with defined yes/no values.

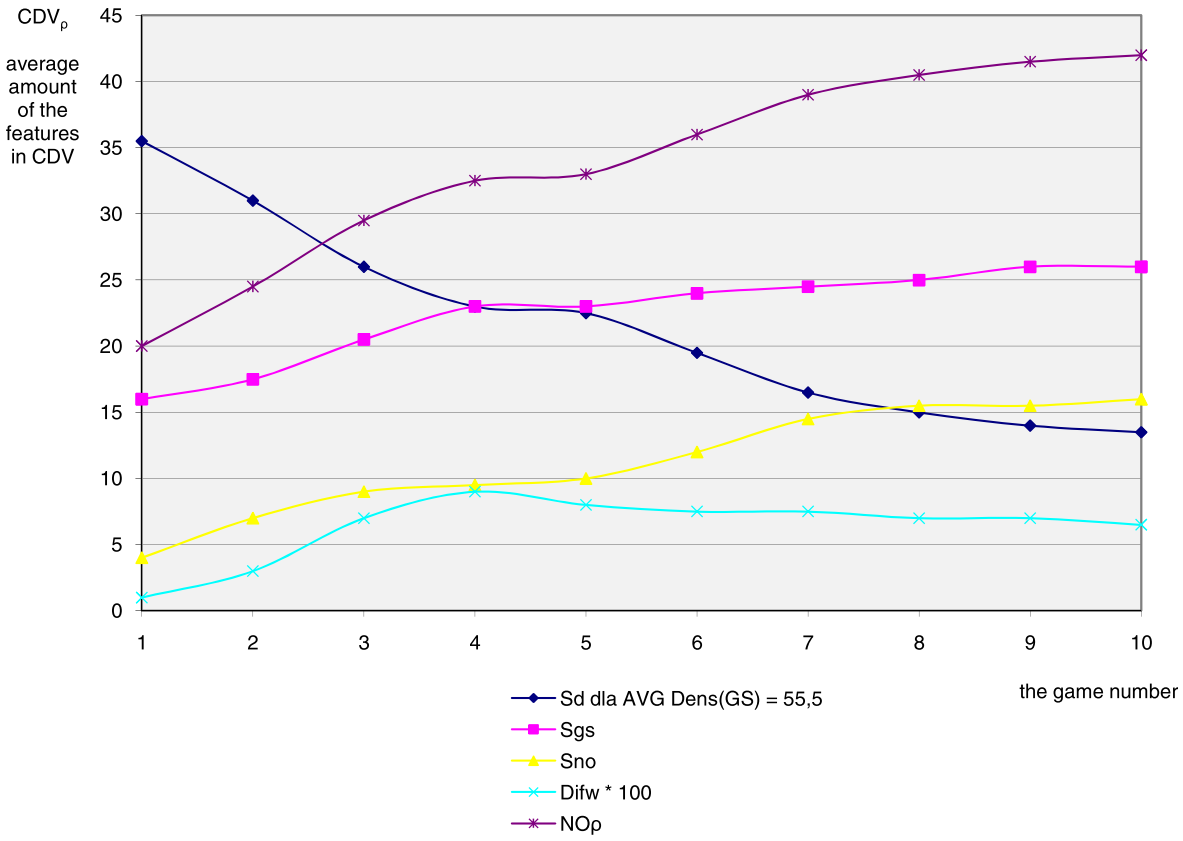

Fig. 7. Dynamics of the process of acquiring new features; averaged results for five new objects.

The S_{GS} value is the number of features from O that are found in the golden standard GS vectors; the reverse measure (S_{NO}) is defined below. The ratio S_d/S_{GS} of the similarity and incompleteness measure shows the percentage of all features of the golden standard that has already been defined for the concept O.

3. $Dif_w = \sum_{i=1}^{m}(|CDV_i(O) - CDV_i(GS)|/m$ is the average difference for all m feature values that appear in both O and GS representations. This measure shows how the feature values differ in O and GS vectors for those features that are common to the two vectors. It allows for observing wrong values associated with relations, while the previous measures allowed only to analyze existence of the relations.
4. $S_{NO} = \sum_{i=1}^{N_f(GS)}[1 - \delta(CDV_i(O) - CDV_i(GS))]$, analogically to S_{GS}, is the measure of similarity of two CDV vectors based on co-occurring features, with summation running over features in the GS. The S_{NO} value is equal to the number of features that appear in description of the concept O and are not found in the GS, thus it measures completeness of the Golden Standard.

The difference between $CDV(O)$ and GS representations is not only due to the lack of knowledge, but also the mechanism to randomize questions, allowing for more knowledge acquisition when the game with the same concept is repeated several time. Results shown in Fig. 7 prove the usefulness of active dialogs for acquiring new knowledge. This can be seen analyzing the graph NO_ρ, ($NO_\rho = S_{NO} + S_{GS}$) where an average increase of the number of features defined in CDV is shown. It can also be observed that during subsequent games the number of the features acquired to the NO grows (graph S_{GS}). For the five test objects the average number of games required to make the system recognize new object correctly was only $V_n = 2.67$. It means that after searching approximately three times for the unknown object the system can identify it correctly. It is also important to notice that the learning process makes objects stable – after the first successful search the object was always correctly recognized in the next games.

S_d value is calculated for the average number of features in all GS which had $Dens(GS) = 55.5$. The decreasing trend of S_d indicates that the number of features in NO comes near to the number of features in GS. It can be expected that after playing more games this value could go below zero indicating that NO has more features than GS. This shows that using the active dialogs one can build better CDV than provided in the Golden Standard which is imperfect, as shown by the S_{NO} graph. The limitations of the GS come from the fact that for the semantic space of 475 features it is hard to acquire full description in the CDV form, even for limited set of objects. Differences (D_{ifw} values) come mostly from the w weights that are negative, making 96.2% of all features in CDVs. It implies that GS is well defined in terms of features that are positively related to the object.

Only a few active dialog templates have been shown here to demonstrate the ability for acquiring common sense knowledge about language stored in the form of weighted vwORF triples. More templates may be added which should lead to improved acquisition of structured knowledge through supervised dialogs in natural language with humans.

7. Comparison to human performance

Results of the semantic search algorithm applied to the 20-questions game may be directly compared to the same task performed by people. In the restricted knowledge domain that our program is working comparison is done between the average number of questions needed to find concepts in games between two people, and in games where one of the players is replaced by the computer program. Let N_q be the number of questions used for guessing an object in the game. Let N_u be the number of unsuccessful searches, so that for N_g games N_u/N_g measures the precision of retrieval.

Experiments were done separately with 4 groups of people of roughly the same size (20–23), or a total of 86 people. First they have been asked to play in pairs the game of questions restricted to the animal domain. In this part of the experiment 93 games have been completed. 86 of these games have been finished in no more than 20 questions and could be used for evaluation. The average number of questions N_q asked by humans to find the object the opponent is thinking about is presented in Fig. 8, with bars labeling groups 1–4, summary of results for games played only by people, and summary of results for games played by people with our program. The height of the bar denotes in logarithmic scale the number of performed games. For each group the minimum and maximum number of questions, the average value and the standard deviation has been presented. Shorter and darker bars represent the number of unsuccessful games, requiring more than 20 questions or missing the target object. Summary of results shows that only a small number of games have been unsuccessful.

In the second phase of the experiment people were asked to perform the game of questions with the computer. Semantic search algorithm with the guessing heuristic (8) has been used. The quality of data stored in the semantic network is estimated by the fraction $Q = 1 - N_u/N_q$ of unsuccessful runs and by the average number of questions N_q required to guess the object, shown in Fig. 8 in the "algorithm" bars.

Knowledge base used in this experiment had 197 objects described by 529 features, with the average number of 50.64 features per object. 227 games performed with people gave the average $Q = 0.64$, which is a bit worse than the results obtained during tests on Semantic Network acquired in an automatic way. This is due to the fact that human players introduced 46 new objects to the ζ knowledge base that had to be learned by the system. Of course 197 objects chosen for automatic network construction does not cover all of the animals domain, but allows for relatively frequent opportunity to learn new objects. If these 46 cases are not included in the number of failed searches much higher quality $Q = 0.81$ is obtained. Searching for new objects caused 56% of all errors, and through the use of active dialogs new knowledge is added and the competence of the system grows.

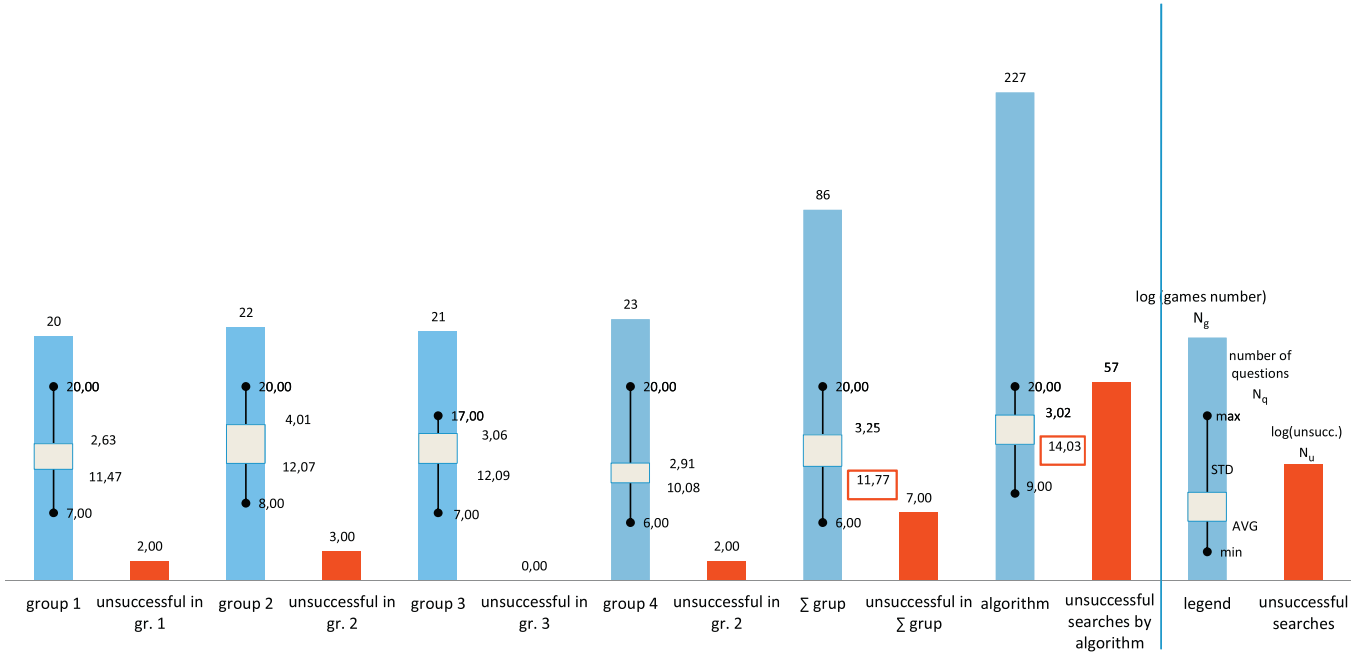

Fig. 8. Comparison of the semantic search algorithm for the 20-question games played in four groups of humans. The bars present: total number of the games played in each of the group as well average results (denoted with $\sum$), number of searches performed by algorithm and number of unsuccessful searches.

The difference between human and semantic search algorithm performance is not that big, on average people used nearly 12 questions to make a correct guess while our algorithm required about 14 questions. The main factor responsible for this difference is probably due to relatively poor quality of knowledge acquired through the automatic data acquisition. However, overall approximation of associative inference mechanisms operating on semantic memory by such simple knowledge representation and search algorithm is quite remarkable. Of course the task itself requires very limited linguistic competences. In most tasks involving natural language processing people use a wide range of common sense knowledge. It seems impossible to obtain such knowledge only from statistical analysis of unstructured text corpora, or even from structured resources such as machine readable dictionaries. What is needed is the active cognition aspect – functionality that allows to verify obtained knowledge in action, which is reduced here to the interaction of the program with people in word games. Introducing more human interactions into the process of lexical knowledge acquisition seems to be a necessary to increase natural language competence of computer systems.

8. Discussion and future research

Cognitive processes rely on different types of memory: recognition, semantic, episodic, working, and procedural memories. We have focused here on the semantic memory as the basis for understanding the meaning of general concepts. Semantic memory as an element of the human cognitive processes has been the subject of many psycholinguistic theories. They are a rich source of inspirations for computational models approximating mental processes used by the brain in language comprehension and production. Such computational models, beside the algorithm, require also a lot of data to operate on. This data should represent knowledge about language concepts and the common sense associations between lexical concepts and their properties. Such lexical data is linked to perception and action in embodied cognitive systems (Ansorge et al., 2010) and thus cannot be easily represented without sensory percepts, their categories and more abstract constructions. At present it is very hard to acquire because it can only be produced and verified by humans, although in future robotic experiments may also offer interesting inspirations. Only a part of symbolic knowledge involved in the description of natural objects and categories may to some degree be captured in semantic networks and create sufficiently rich relations to grant symbols some elementary meaning. Such representation of knowledge allows for approximation of natural processes responsible for language comprehension.

In this paper a step towards computationally efficient model of semantic memory has been made. Knowledge representation in the form of weighted triples *vwORF*, has been used for implementation of functionalities inspired by psycholinguistic theories of human semantic memory. Semantic network build from the *vwORF* atoms of knowledge is a flexible way of storing lexical knowledge. For computational efficiency vector-based semantic space is used instead of semantic network. Elementary linguistic competence has been demonstrated in word games in this way, with results that have not been shown so far by more sophisticated linguistic approaches. Word games are a good ground for comparison of human competencies with capabilities of computational models. Results achieved by people and by our semantic search approach based on *vwORF* knowledge representation in the 20-question game show that although real brains are still better than computer programs the difference is not so large.

Linguistic competence of programs depends more on lexical knowledge, representation scheme and search algorithm than on raw computational power. Research on expert systems showed the difficulty and the importance of knowledge acquisition from data, and despite the availability of huge structured and unstructured lexical resources acquiring lexical information automatically is still a great challenge. Using three independent lexical databases we have shown possibilities for obtaining automatically common sense knowledge in the form of typed relations between lexical elements. Nevertheless, the data need to be validated and corrected interacting with people. Active dialogs introduced here allow for acquisition of common sense knowledge and verification of this knowledge in action. This is frequently omitted in construction of large lexical databases, such as WordNet that has been built manually with a great effort. Without systematic feedback from active use of its resources the process of completing missing knowledge and proper stratification of Wordnet synsets in different contexts is very slow.

Semantic search process introduced in this paper may be treated as a general model of decision making based on active queries, to find particular action (object) appropriate in specific conditions (feature values). Consider for example the process of medical diagnosis where disease is identified using a series of observations and tests; decision support system should ask a number of questions to identify the most distinctive symptoms. The algorithm has already been tested in medical domain using data from the "Diagnostic and Statistical Manual of Mental Disorders" (DSM IV) (DSM, 1994). Queries generated by semantic search led to correct diagnosis in fewer steps than the original decision tree recommended by DSM IV. Other applications include WWW information retrieval (Duch & Szymański, 2008). Web search engines return a large set of pages as a result of keyword-based query, and the subset of most relevant pages is subsequently identified using the semantic search algorithm. However, such approach requires features relevant for concepts contained in all possible knowledge domains indexed by the search engine. It implies building a very large scale semantic network which is still a great challenge. This *vwORF* representation of knowledge has also been successfully applied for

improvement of natural language text processing (Majewski & Szymański, 2008).

The tests performed here in the limited domain should be treated as the proof of concept. The project will be scaled up and used to improve information retrieval from Wikipedia. Interaction of many volunteer contributors would be needed to create knowledge for a large scale semantic network, verified in action during the actual searches. A good strategy is to start form a limited domain, such as animals or plants, trying to cover the whole domain, not just a small subsets and has been done here. Identifying an arbitrary plan or animal shown in a photograph using a variant of the 20-question game is a challenging task. Going beyond simple nouns and trying to understand actions is still farther ahead. In all these tasks neurocognitive inspirations should be our guide.

Acknowledgment

This work has been supported by the Polish Committee for Scientific Research Grant N516 035 31/3499.

References

Ansorge, U., Kiefer, M., Khalid, S., Grassl, S., & Knig, P. (2010). Testing the theory of embodied cognition with subliminal words. *Cognition, 116*, 303–320.

Buckland, M., & Gey, F. (1994). The relationship between recall and precision. *Journal of the American Society for Information Science, 45*(1), 12–19.

Burke, D., MacKay, D., Worthley, J., & Wade, E. (1991). On the tip of the tongue: What causes word finding failures in young and older adults. *Journal of Memory and Language, 30*(5), 542–579.

Collins, A., & Loftus, E. (1975). A spreading-activation theory of semantic processing. *Psychological Review, 82*(6), 407–428.

Collins, A., & Quillian, M. (1969). Retrieval time from semantic memory. *Journal of Verbal Learning and Verbal Behaviour, 8*, 240–247.

Cullen, J., & Bryman, A. (1988). The knowledge acquisition bottleneck: Time for reassessment?. *Expert Systems 5*(3), 216–225.

Davis, R., Shrobe, H., & Szolovits, P. (1993). What is a knowledge representation? *AI Magazine, 14*(1), 17–33.

DSM (1994). Diagnostic and statistical manual of mental disorders. American Psychiatric Association.

Duch, W. (2007). What is computational intelligence and where is it going. In W. Duch & J. Mandziuk (Eds.). *Challenges for computational intelligence* (Vol. 63, pp. 1–13). Springer.

Duch, W., Matykiewicz, P., & Pestian, J. (2008). Neurolinguistic approach to natural language processing with applications to medical text analysis. *Neural Networks, 21*(10), 1500–1510.

Duch, W., & Szymański, J. (2008). Semantic web: Asking the right questions. In *Proceedings of the 7 International Conference on Information and Management Sciences* (pp. 1–8). California Polytechnic State University.

Feldman, J. A. (2006). From molecule to metaphor: A neural theory of language. MIT Press.

Forster, K. (2004). Category size effects revisited: Frequency and masked priming effects in semantic categorization. *Brain and Language, 90*(1–3), 276–286.

Gleason, J. B., & Ratner, N. B. (1997). Psycholinguistics (2nd ed.). Wadsworth Publishing.

Goldberg, D. (1989). *Genetic algorithms in search, optimization and machine learning*. Boston, MA, USA.: Addison-Wesley Longman Publishing Co., Inc.

Guarino, N., & Poli, R. (1995). Formal ontology, conceptual analysis and knowledge representation. *International Journal of Human Computer Studies, 43*(5), 625–640.

Hunston, S. (2001). Word frequencies in written and spoken english: Based on the british national corpus. *Language Awareness, 11*(2), 152–157.

Lamb, S.M. (1999). Pathways of the brain: The neurocognitive basis of language (Vol. 170). John Benjamins Publishing Company.

Liu, H., & Singh, P. (2004). ConceptNet. A practical commonsense reasoning tool-kit. *BT Technology Journal, 22*(4), 211–226.

Majewski, P., & Szymański, J. (2008). Text categorisation with semantic common sense knowledge: First results. bSpringer Lecture notes in computer science. In *Proceedings of 14th int. conference on neural information processing (ICONIP07)* (Vol. 4985, pp. 285–294).

Matykiewicz, P., Pestian, J., Duch, W., & Johnson, N. (2006). Unambiguous concept mapping in radiology reports: Graphs of consistent concepts. *AMIA Annual Symposium Proceedings, 2006*, 1024–1031.

McClelland, J., & Rogers, T. (2003). The parallel distributed processing approach to semantic cognition. *Nature Reviews Neuroscience, 4*(4), 310–322.

McNamara, T. (2005). Semantic priming: Perspectives from memory and word recognition. UK: Psychology Press.

McNamara, T. P. (2005). Semantic priming: Perspectives from memory and word recognition. Taylor & Francis Group: Psychology Press.

Miikkulainen, R. (1993). Subsymbolic natural language processing: An integrated model of scripts, lexicon, and memory. Cambridge, MA: MIT Press.

Miller, G. A., Beckitch, R., Fellbaum, C., Gross, D., & Miller, K. (1993). Introduction to WordNet: An on-line lexical database. Princeton University Press.

Mitchell, T., Shinkareva, S., Carlson, A., Chang, K., Malave, V., Mason, R., et al. (2008). Predicting human brain activity associated with the meanings of nouns. *Science, 320*(58–80), 1191.

Ogden, C., Richards, I., Malinowski, B., & Crookshank, F. (1949). The meaning of meaning. Routledge & Kegan Paul.

Posner, M., & Keele, S. (1968). On the genesis of abstract ideas. *Journal of Experimental Psychology, 77*(3), 353–630.

Pulvermnller, F. (2003). *The neuroscience of language on brain circuits of words and serial order*. Cambridge Uni. Press.

Qian, G., Sural, S., Gu, Y., & Pramanik, S. (2004). Similarity between Euclidean and cosine angle distance for nearest neighbor queries. In *Proceedings of the 2004 ACM symposium on applied computing* (pp. 1232–1237).

Quinlan, J. (1986). Induction of decision trees. *Machine Learning, 1*(1), 81–106.

Rosch, E. (1973). Natural categories. *Cognitive Psychology, 4*(3), 328–350.

Rumelhart, D. E., & McClelland, J. L. (Eds.). (1986). *Parallel distributed processing: Explorations in the microstructure of cognition*. Cambridge, MA: MIT Press.

Smith, E., Shoben, E., & Rips, L. (1974). Structure and process in semantic memory: A featural model for semantic decisions. *Psychological Review, 81*(3), 214–241.

Sowa, J. (1991). *Principles of semantic networks: Explorations in the representation of knowledge*. *Representation and reasoning*. San Mateo, CA: Morgan Kaufmann.

Speer, N. K., Reynolds, J. R., Swallow, K. M., & Zacks, J. M. (2009). Reading stories activates neural representations of perceptual and motor experiences. *Psychological Science, 20*, 989–999.

Staab, S., & Studer, R. (2004). Handbook on ontologies. Springer Verlag.

Szymański, J., Dusza, K., Byczkowski, L. (2007). Cooperative editing approach for building wordnet database. In: *Proceedings of the XVI international conference on system science* (pp. 448–457).

Szymański, J., Sarnatowicz, T., & Duch, W. (2007). Towards avatars with artificial minds: Role of semantic memory. *Journal of Ubiquitous Computing and Intelligence*.

Tulving, E., Bower, G., & Donaldson, W. (1972). Organization of memory. New York: Academic Press.

Vanderwende, L., Kacmarcik, G., Suzuki, H., & Menezes, A. (2005). MindNet: An automatically-created lexical resource. *Proceedings of HLT/EMNLP on Interactive Demonstrations*, 8–19.

Voss, P. (2005). *Essentials of general intelligence: The direct path to artificial general intelligence*. *Artificial general intelligence*. Springer, pp. 131–157.

Warren, R. (1977). Time and the spread of activation in memory. *Journal of Experimental Psychology: Human Learning and Memory, 3*(4), 458–466.

Weizenbaum, J. (1966). ELIZA – A computer program for the study of natural language communication between man and machine. *Communications of the ACM, 9*(1), 36–45.

Zadeh, L. (1996). Fuzzy logic = computing with words. *IEEE Transactions on Fuzzy Systems, 4*(2), 103–111.