Theoretical Computer Science 488 (2013) 2–14

Contents lists available at SciVerse ScienceDirect

Theoretical Computer Science

journal homepage: www.elsevier.com/locate/tcs

Context granulation and subjective-information quantification

Sankar K. Pal, Romi Banerjee *

Center for Soft Computing Research, Indian Statistical Institute, Kolkata 700 108, India

ARTICLE INFO

Article history:
Received 17 June 2012
Received in revised form 2 November 2012
Accepted 19 December 2012
Communicated by A. Skowron

Keywords:
Bayesian inference
Cognition
Computing with words
Concept graph generation
Context dependence
Contextual keywords
Entropy
Granulation
Linguistics
Perceptions
Semantic disambiguation
Text summarization
Z-numbers

ABSTRACT

In the course of comprehension, the human brain makes astounding inferences despite insufficient and ambiguous data—the data mostly being 'words' that frame sentences. Can this neural process of "abstraction of the relevant" be sufficiently modeled such that the computer may be induced to perform the same? This article is our response to this very question; a treatise on our attempt at engineering the basis of such a model, or rather, a methodology of 'relevance cognition'. Drawing inspiration from the way children learn languages, we propose here a corpus and entropy-based methodology that follows a subjectivist approach to granulate and identify statements of consequence in a natural language sample, all within the periphery of a particular context. Besides promoting language grasping abilities of a machine – through simulation of the intuitive process of relevant-information segregation – the methodology aims at reducing overall text-processing costs. The suggested scheme considers 'conceptual keywords' to be the basis of sentential understanding and utilizes the principles of Bayes' conditional probability and Shannon's entropy theorem. Experimental results have been provided to substantiate our logic. Though this article has been formulated over the backdrop of the Z-number approach to computing with words (CWW), it nevertheless applies to research areas in natural language processing (NLP) like text summarization, semantic disambiguation and concept graph generation.

1. Introduction

> "*How do our minds get so much from so little? We build rich causal models, make strong generalizations, and construct powerful abstractions, whereas the input data are sparse, noisy, and ambiguous—in every way far too limited. A massive mismatch looms between the information coming in through our senses and the outputs of cognition* [1]".

Given a text or a speech sample, cognitive language comprehension proceeds in the following sequence [2]:

(a) Obtain a primitive abstract concept from textual sentences – identify the 'context' of discourse and consequent interpretation of the sentences;
(b) Compute the weights of logical relations among sentences according to cognitive principles—group sentences into logical granules or subcontexts;
(c) Measure the textual context in sentential granularity based on the weights—identify the granules, or rather the sentences that are significant.

* Corresponding author.
E-mail addresses: sankar@isical.ac.in (S.K. Pal), rm.banerjee@gmail.com (R. Banerjee).

doi:10.1016/j.tcs.2012.12.034

Each of the above steps depends on the 'words' that constitute the sentences in the text. Words being inherently uncertain – they have context-sensitive interpretations, the adjectives and adverbs convey different perceptions across individuals, have multiple syntactic or morphological forms, and may not be used in their literal sense – the questions that challenge scientists who wish to model intuitive text-conceptualization are, how does the brain:

(a) Overcome the word-perception ambiguities to identify contexts and subcontexts?
(b) Discard irregularities to extract all significant sentences?
(c) Integrate the meanings of these important sentences into cognitive granules that summarize the entire text?

> "*When children learn their first words, they face a challenging joint-inference problem*:
>
> (a) *They are trying to infer what meaning a speaker is attempting to communicate at the moment a sentence is uttered; and*
> (b) *Trying to learn the more stable mappings between words and referents that constitute the lexicon of their language.*
>
> *With either of these pieces of information, their task becomes considerably easier. Knowing the meanings of some words, a child can often figure out what a speaker is talking about, and inferring the meaning of a speaker's utterance allows a child to work backward and learn basic-level object names with relative ease. However, for a learner without either of these pieces of information, word learning is indeed a hard computational problem. Adults face the challenge of learning entirely novel object concepts less often* [3]".

The above excerpt leads us to understand that the instinctive process of learning new words and their meanings by children proceeds as follows:

(a) A child (C) is initially taught the meanings of a few basic words;
(b) When C reads a sentence (S), it tries interpreting S on the basis of its current vocabulary. In the case of a failure, C refers to a dictionary or asks someone.
(c) These new words are accordingly associated with appropriate existing words in C's vocabulary, such that it is able to use the new word(s) correctly the next time.

    This is synonymous to C's identifying and learning new words.
(d) Steps (b) and (c) are continual processes.

If one were to deliberate over the above points, it should be quite evident that capacitating the computer to compute with words is indeed, hardly any different from the process of learning words and their usage by children.

A machine is devoid of natural cognitive abilities. Thus, if by any means, a machine inculcates the ability to identify and learn new words, it should be able to comprehend any given language sample, at least as well as children. This is exactly what we aim at in the research described in this article—modeling the process of intuitive language understanding, based on all the concepts mentioned above. The research underlying this article, though is spurred by the need to reduce the computational complexities inherent in our algorithm [4] for the Z-number approach to computing with words (CWW), aims to achieve a lot more, thereby contributing to the field of CWW, natural language processing (NLP) and artificial intelligence in general.

Motivated by the natural process of information segregation in the course of text comprehension, we propose here, a subjectivist corpus-based methodology to granulate and rank sentences (S) on the basis of the importance of their intrinsic information (visualized as the entropy of S). Not only would such an operation address the language comprehension issue but would also lead to the elimination of all irrelevant text granules, thereby reducing text-processing costs. The proposed methodology focuses on the identification of keywords in a sentence as the agent of comprehension and is based on the principles of Bayes' theorem of conditional probability and Shannon's entropy theorem.

Drawing on the philosophy of word-learning by children, our work involves the creation and the maintenance of a word-corpus—a list of keywords and key-phrases, pertaining to a context. Moreover, considering the enormous existing and continually growing vocabulary, be it any context—health services, banking services or literature in any genre, and that most words have multiple perceptions (homonyms, polysemes, heteronyms, capitonyms), we believe it would be pragmatic to base our work on a specific context-sensitive interpretation of words. The context we have considered is 'mystery' stories, particularly written by Agatha Christie.

The words in our corpus are augmented with: (a) the subcontext and (b) the frequency of occurrence of the word in the subcontext across texts experienced. Subcontexts imply finer subdivisions within the macro-context (as for instance, if 'banking services' was the context, 'fixed-deposit issues' could be a possible subcontext), while the frequency is proportional to the probability of the word being used or applicability of the word to the subcontext and the context in turn. The word-corpus serves as the machine's lexicon, and is updated every time a new keyword (incorporation of the new word) or an existing keyword (updating of its subcontextual frequency) is found. The frequency of occurrence is synonymous to a measure of the familiarity with the word, i.e., the greater the frequency, the better the machine is acquainted with the words. The merits of a corpus-based study can be found in [5].

Simulation of the concepts mentioned above presents the need to evaluate the information content in a sentence. [6–10] are significant treatises on the variations of entropy in a text sample. These articles illuminate upon the distribution of unconditional and conditional information and the decay of entropy in texts. [11] uses the lexical chaining method to granulate sentences and the vector spaces approach to evaluate sentences of importance in a text sample, leading to text summarization. While none of these articles attempt at modeling the cognitive text comprehension process, this paper endeavors to do so.

> *"It must be recognized that the notion of "probability of a sentence" is an entirely useless one, under any known interpretation of this term."*—Noam Chomsky (1969)

Previous studies on the evaluation of the entropy of a sentence (S) were based on the information in S being the negative logarithm of the 'probability' of S. These studies primarily depended on the n-gram model [12] of text representation, where the probability of a sentence is the product of the probabilities of its constituent words, and the probability of each word (w) is conditioned on the words that occur before w in the sentence. Understandably, methods based on the 'probability of a sentence' approach differ radically from Chomsky's line of thought.

We, however, deem Chomsky's words invaluable. Evaluating the probability of some S requires analyzing large volumes of corpus data to measure probabilities of words and word-associations. Considering the numerous ways a single expression (E) may be worded, it practically is impossible to collect or acquire a 'complete' list of all possible word sequences meaning a given E. Moreover, the probability of usage of sentences – particular to some context – is unique to the linguistic abilities of the author. Thus, our study deviates from the n-gram model and is based on the consideration of 'keywords' as the element in S that guides comprehension—analogous to the way children acquire linguistic skills.

In this article, we do not deal with the complexities involved in differentiating between the types of sentences nor with the evaluation of information conveyed by sentence granules. What we wish to model is the information that a reader or the receiver of the information is able to extract, given a text sample, and not the information that the sender might want to convey.

The article begins with an acknowledgment of the topics that inspired such an investigation (Section 2), followed by brief notes on the concepts underlying the proposed theory (Section 3). The paper then moves on to a description of our formulated methodology (Section 4) to end with a depiction of the results received and directions of future research in the area (Section 5). Section 5 also has a note on the data structure designed to maintain our corpus.

2. The inspirations

2.1. Computing with words (CWW) and the Z-numbers

Breaking away from the traditional methods of computation using numbers, CWW stands for computation using words and phrases in natural language statements. Coined [13] by Zadeh, the paradigm acknowledges the immense cognitive ability of the human brain to formulate decisions on the basis of perceptions of words framing natural language statements. These perceptions may be ambiguous, incorrect or even biased.

Human beings intuitively 'compute' with words when:

(a) The situation lacks numeric precision;
(b) The situational ambiguity may be exploited to arrive at low-cost, robust, tractable results;
(c) Words express a lot more than numbers;
(d) The problem cannot be solved by numeric computation principles.

Thus, CWW aspires to capacitate the computer to learn, understand and use words (basically adjectives, adverbs and figures of speech) the way human beings do. Consequently, it might be stated that if a machine were able to compute with words, it would actually possess a measurable amount of intelligence, i.e., a machine intelligence quotient (MIQ [14]).

A system [4] based on CWW, takes as input a set (S) of natural language statements, converts S into some precisiated machine understandable form (S') based on the constituent 'words', and processes S' to arrive at resultant keywords which are then framed into natural language responses. The generic architecture of such a system can be intuitively defined as shown in Fig. 1. The architecture comprises the following units:

(a) The Encoder—Translates the input statements into precisiated antecedent constraints;
(b) The Rule Base—Consists of context-sensitive antecedent–consequent event relationships. For our studies, we assumed the Rule Base to be a collection of antecedent–consequent event relationships as well as commonsense primitives [15]. The rule base is in the form of an Explanatory Database (ED);
(c) The Inference Engine, in conjunction with the rule basis leads to the input antecedent constraints being processed to arrive at the appropriate consequents. These results are in some symbolic form;
(d) The Decoder—Translates the symbolic outputs into semantically correct natural language statements.

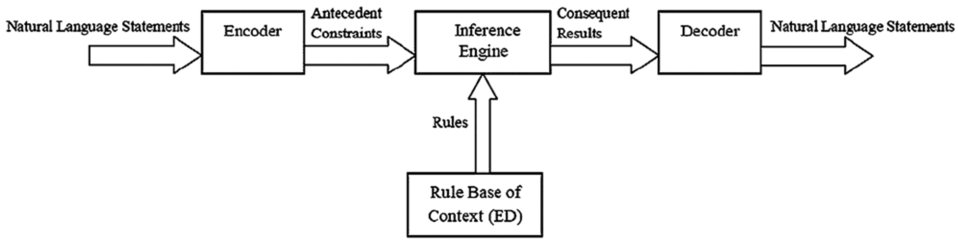


Fig. 1. The generic architecture of a system based on CWW [4].

The principles of CWW, as envisioned by Zadeh, are rooted in [13,16–19]; and over the last decade, there has been substantial research on different areas that underlie the paradigm. [20–27] provide valuable insights into the scientific investigations done in this domain.

The Z-number [28] is a novel fuzzy-theoretic approach to CWW, proposed by Zadeh in 2011, and the concept serves as a model for the precisiation of the perception(s) embedded in a natural language statement. The Z-number methodology is where our focus lies.

Given a natural language statement (S), with a subject (X), the corresponding Z-number is a 2-tuple: $\langle$value (A) of X in S, certainty (B) of $A\rangle$. The parameter B is an indication of the perceived certainty or reliability of the information conveyed by S, i.e., it includes an element defining human affects associated with the certainty-perception of A. X is fundamentally considered to be a linguistic variable with A as its linguistic value; both A and B are typically represented as trapezoidal fuzzy numbers.

The ingenuity, algorithms and subsequent implementation challenges of the Z-number have been thoroughly investigated in [4]. All discussions in this article arise from our algorithm for CWW using the Z-number approach. While our algorithm views the Z-number as an aid to NLP and seeks a seamless union of CWW and NLP, it involves the following complexities per sentence (S) in a text sample:

(a) Test for S's relevance to the context;
(b) Identification of S's sentence type—simple, complex, compound, declarative, interrogative, exclamatory, imperative;
(c) Decomposition of S into its simple sentence set (S'), if S is compound or complex;
(d) Extraction of the Z-valuation parameters X, A and B, per element of S';
(e) Assimilation of the constituent Z-valuations into a complete mathematical expression (E) denoting S;
(f) Evaluation of E, resulting in some Z-valuation(s);
(g) Translation of Z-valuations into semantically correct natural language sentences.

Areas (a) through (e) fall under the jurisdiction of the encoder, while (f) and (g) under that of the inference engine and the decoder, respectively; and understandably, factors (b) through (g) come into consideration only if S is context relevant. Thus, if the impact or relevance of information conveyed by each S is measured, prior to the extraction of the Z-numbers, irrelevant sentences need not be processed—leading to a significant reduction in general text processing costs. The algorithm formulated in this paper seeks to resolve this relevance-cognition issue.

2.2. Concepts of communication and information theory

Shannon, in his celebrated 1948 paper [29], puts forth the following essential concepts on communication and information theory:

(a) "The fundamental problem of communication is to reproduce at one point either exactly or approximately a message selected at another point. Frequently the messages have *meaning*; that is they refer to or are correlated according to some system with certain physical or conceptual entities".
(b) The entropy of a message is a measure of the information it conveys. Considering each independent component of the message to be a random variable, the entropy of the message is the sum of the uncertainty associated with the value of the random variables in the message.
(c) The noisy-channel coding theorem implies that transferring data at a constant rate is the most efficient way of communicating information across noisy channels—the theoretical upper bound of the transfer rate being given by the Shannon capacity, for a given noise level.
(d) English speech functions with the probability measure given by the frequency of occurrence in ordinary use.

These concepts make it apparent that real communication through language does indeed abide by the above facts. Communication can only be successful if there is a sufficient overlap between the concept conveyed and that received. This is possible if there is parity between the rate of transfer of information and the rate of comprehension of the receiver. The random variables in a message imply the words that constitute the message. These concepts are pivotal to the logic presented in this paper.

The following section comprises of short notes on facts that underlie the formulated theory.

3. Building up the foundations of the proposed algorithm

3.1. Context and context-relevance

Languages are a means of 'communicating' thoughts and ideas. Every language has its own alphabet – basic symbols – that are combined to form words. Words are independent units of any language that are used pragmatically, semantically or conceptually to express something and word-interpretations are functions of the nature of their usage, i.e., the morphological form and the context. Natural language is thus intrinsically 'context sensitive'.

The following excerpt from [30], perfectly summarizes the context-sensitivity of natural language, and is a clear indication of what a text-processing system is required to accomplish. The proposed algorithm is unquestionably far from realizing every aspect but can surely be considered a step towards seeing it achieved.

> *"If all the words in a language had unique and determinate meanings (no ambiguity or vagueness) and fixed references (no indexicality), and if using language were simply a matter of putting one's thoughts into words, understanding an utterance would merely be a matter of deciphering whatever words the speaker uttered. But language and our use of it to communicate are not as straightforward as that. Some expressions, most obviously pronouns, like 'I', 'they', and 'this', and temporal terms like 'today' and 'next week', do not have fixed references. For example, when I use 'I' it refers to me, but when you use it it refers to you. Moreover, we often speak inexplicitly, non-literally, or indirectly, and in each case what we mean is distinct from what can be predicted from the meanings of the expressions we utter. We can leave something out but still mean it, use a word or phrase figuratively, or mean something in addition to what we say. We even can do all three at once."*

[31–37] offer a deeper insight into the ambiguities of natural language.

A 'context' implies a perception granule, a cognitive unit [2] that helps identify and model a real-world or abstract entity. A context is implied by key nouns or noun-phrases in a sentence, textual semantic relations—between individual words and sentences, common sense, knowledge [38], communication background etc. An object is 'context relevant' if it bears an identifiable relation to the context of discussion.

Linguists claim that contexts are crucial to the comprehension of a concept. Authors in [39] illustrate the validity of the statement using the concept of Boolean function minimization.

Context relevance can be categorized into:

(a) Intra-relevance: relevance of a single sentence to the topic of discourse;
(b) Inter-relevance: relevance of sentence across a text sample. This is the factor which allows grouping of sentences into granules of subcontexts.

The proposed algorithm tries to address both these types of relevance.

3.2. Bayes' Theorem

Bayes' Theorem, originally stated by the Reverend Thomas Bayes, is visualized as a method of understanding the effect of a new piece of evidence on the truth of a theory [40]. The success of the Bayesian approach towards understanding the process of cognitive development has inspired powerful and humanlike approaches to machine learning for the past two decades. It forms the cornerstone of the *subjectivist* approach to epistemology, statistics, and inductive logic.

Subjectivists profess 'learning' to be a continual process of revision of 'belief' in which a 'prior' subjective probability (P) is replaced by a 'posterior' probability (Q), the latter incorporating newly acquired information [41]. Given an event ($\underline{E}$), the 'learning' process is believed to proceed in the following sequence:

(a) Some of the probabilities associated with sub-events underlying E, are altered by experience, intuition, memory, or some other non-inferential learning process;
(b) The rest of the opinions are updated to bring them in line with the newly acquired knowledge.

Derivation of Bayes' formula [42]**:**

Let E and F be events, where E may be expressed as,

$$E = EF \cup EF^c \tag{1}$$

where, F^c stands for the complement of F, and EF and EF^c denote intersection of the events E and F and E and EF^c respectively.

The conditional probability that E occurs given that F has occurred is denoted as $\Pr(E|F)$, such that,

$$\Pr(E|F) = \frac{\Pr(EF)}{\Pr(F)}. \tag{2}$$

Since, EF and EF^c are mutually exclusive,

$$\begin{aligned}\Pr(E) &= \Pr(EF) + \Pr\left(EF^c\right)\\ &= \Pr(E|F)\Pr(F) + \Pr\left(E|F^c\right)\Pr\left(F^c\right)\\ &= \Pr(E|F)\Pr(F) + \Pr\left(E|F^c\right)(1-\Pr(F)).\end{aligned} \tag{3}$$

Equation (3) may be generalized as follows,
Let, $F_1, F_2, \ldots, F_n$ be mutually exclusive events, then,

$$E = \bigcup_{i=1}^{n} EF_i \tag{4}$$

$$\begin{aligned}\Pr(E) &= \sum_{i=1}^{n} \Pr(EF_i)\\ &= \sum_{i=1}^{n} \Pr(E|F_i)\Pr(F_i).\end{aligned} \tag{5}$$

Now, using (2) and (5), we arrive at Bayes' formula,

$$\begin{aligned}\Pr\left(F_j|E\right) &= \frac{\Pr(F_jE)}{\Pr(E)}\\ &= \frac{\Pr\left(E|F_j\right)\ \Pr(F_j)}{\sum_{i=1}^{n}\Pr(E|F_i)\Pr(F_i)}.\end{aligned} \tag{6}$$

Bayes' theorem forms the logic underlying the updating of our word-corpus.

3.3. Shannon's entropy

In [29], Shannon attempts to quantify information in a message and introduces the concept of entropy or the 'uncertainty' of information. This entropy is attributed to the probability of occurrence of the symbols constituting the message—the greater the probability of occurrence, the lower is the 'information' content of the symbol.
Following Shannon's theorem of entropy,
Let there be a set of n possible events, where $p_1, p_2, \ldots, p_n$ denote the probability of occurrence of each of the events, respectively, such that,

$$\sum_{i=1}^{n} p_i = 1. \tag{7}$$

If all of the events are equally probable,

$$p_i = \frac{1}{n}, \quad \forall i,\ i = 1,\ 2, \ldots n. \tag{8}$$

The information (I) conveyed per event is:

$$I_i = -\log p_i. \tag{9}$$

The expectation (E) of information per event:

$$E_i = -p_i \log p_i \tag{10}$$

where, i denotes the event and $i = 1, 2, \ldots, n$ and the base of the logarithm depends on the radix of the coding system.
Thus, if X is a discrete random variable on a finite set, $\mathcal{X} = \{x_1,\ x_2, \ldots,\ x_n\}$ with probability distribution $p(x) = \Pr(X = x)$, the entropy (average information over all the events) $H(X)$ of X is defined as,

$$H(X) = -\sum_{x \in \mathcal{X}} p(x) \log p(x). \tag{11}$$

$H(X)$ is often denoted as $H_n(p_1,\ p_2, \ldots, p_n)$ where p_i denotes the probability $Pr(X = x_i)$.
Some of the basic properties of Shannon's entropy are:

(a) For any n, $H_n(p_1,\ p_2, \ldots, p_n)$ is continuous and symmetric on $p_1, p_2, \ldots, p_n$;
(b) An event with zero probability, does not contribute to the entropy;
(c) The entropy is maximized when the probability distribution is uniform.

Shannon's theorem forms the basis of the measure of information of sentences in our algorithm.
Following this brief overview of the underlying concepts, the article now moves on to the proposed logic.

4. The proposed logic

4.1. Formulation of the entropy of a sentence based on the principles of Bayesian inference and Shannon's entropy theorem

Considering each natural language statement (S) in a text sample to be a 'message' and the constituent keywords to be random events,
Let,

(a) C be some context of discourse;
(b) $C = \{c_1, c_2, \ldots, c_n\}$, $c_i = i$th subcontext under C;
(c) $\Pr(c_i)$ = prior probability of occurrence of c_i = percentage of occurrence of c_i across all subcontexts experienced = familiarity of the machine with c_i;
(d) The corpus of C be defined by a set of 3-tuples:

$\langle$keyword (w), frequency ($f_{w\_c_i}$) of occurrence of w in c_i, $c_i\rangle$;

$$\text{where,}\quad f_{w\_c_i} = \frac{No.of\ \ texts\ which\ contain\ w\ used\ in\ the\ subcontext\ c_i}{Total\ number\ of\ \ texts\ faced} \tag{12}$$

and, $f_{w\_c_i}$ = familiarity of the machine with w in c_i.

(e) A single keyword might occur in various subcontexts, with a different probability of occurrence per subcontext;
(f) $|S|$= number of keywords in S .

Note:

Existing corpus definitions, take into account the total intra-document frequency of occurrence per term. We, however, use the inter-document frequency to define the corpus; reasons being twofold –

(a) The corpus does not require to be updated at every intra-document occurrence of the word;
(b) The intra-document frequency is an indication of the word being a 'keyword' with respect to the document, while the inter-document frequency should determine if the word is a 'keyword' with respect to the given context.

Therefore,

(a) From the principles of Bayesian inference:

The probability of occurrence of w, given c_i is,

$$\Pr\,(w|c_i) = f_{w\_c_i} \tag{13}$$

and, using (6) and (13), the posterior probability of c_i given a word w in a sentence S is,

$$\Pr\,(c_i|w) = \frac{\Pr\,(w|c_i)\Pr(c_i)}{\sum_{j=1}^{n}\Pr(w|c_j)\Pr(c_j)}. \tag{14}$$

(b) From the principles of Shannon's entropy:

Following (9) and (14), the information conveyed by w, within the subcontext c_i, can thus be defined as,

$$I_{w\_c_i} = -\log_{26}\Pr\,(c_i|w) \tag{15}$$

and, therefore from (11) and (15), the entropy of the entire sentence S, that lies in the subcontext c_i is,

$$\mathrm{H}\,(S) = -\frac{1}{|S|}\sum_{j=1}^{|S|}\Pr\left(c_i|w_j\right)\log_{26}\Pr\left(c_i|w_j\right) \tag{16}$$

where, $w_j = j$th keyword in S.

The base of the logarithm operations in (15) and (16) is 26 as the English alphabet has 26 characters. The normalizing factor $\left(\frac{1}{|S|}\right)$ in (16) averages out the quantity of information across the total number of keywords present in the sentence, thus negating the bias towards sentences with a larger number of keywords.

The following sub-section presents our formulated algorithm, which uses (16) to annotate each sentence (S) in a language sample by its entropy. The greater the entropy of S, the greater is its information content and the greater is its relevance in comprehension of the text.

4.2. Algorithm — recognition of the granules of subcontext of discourse and identification of the consequential sentences in a text sample

Assumptions about the machine (M):

M is trained:

(a) To annotate the words in the text with their respective parts of speech (pos) [43];
(b) To map pronouns to the corresponding nouns in the text-sample [44];
(c) In identifying the subcontexts that fall under the context considered, by virtue of commonsense primitives [15];
(d) To identify keywords and associate them with subcontexts using WordNet [45,46];
(e) To update its text-corpus (TC)—add new words and update frequencies of occurrence of existing words.

Assumptions about the text sample (T):

(a) All the words are used in their literal sense, i.e., the words are taken at face value;
(b) All the sentences are syntactically and semantically correct;
(c) Each sentence belongs to a single subcontext;
(d) There exists a semantic coherence in the sentences within a paragraph, i.e. each paragraph pertains to a single line-of thought or subcontext—as is expected in all well-written texts.

Input:

(a) The text sample (T).
(b) The initial text-corpus (TC).

Output:

Granules of subcontexts containing sentences ranked in descending order on the basis of their entropy value.

Notation conventions:

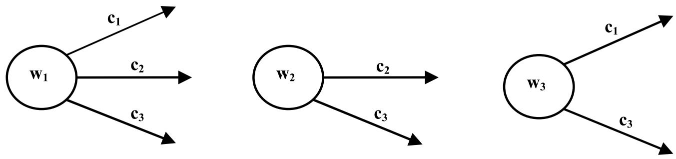
(a) $P =$ paragraph $\in T$.
(b) $S =$ sentence under observation, $\exists P : S \in P \in T$.
(c) $S\_no =$ position of S in P
(d) $w =$ keyword in S.
(e) $c =$ the subcontext of S.

Steps:

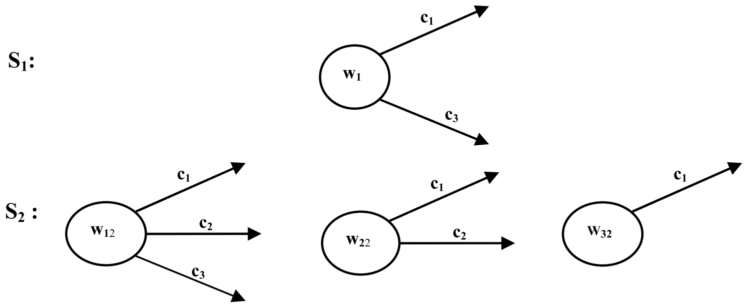
1. For all $P \in T$
   Repeat steps 2 through 11
2. Annotate words with their parts of speech
3. Map all pronouns to their corresponding nouns
4. Remove all stop-words
5. Extract keywords
6. Identify c that is common to all the keywords
7. Update TC (based on the identified c, $\Pr(c)$ and $\Pr(c\,|w)$)
8. For all $S \in P$
   Repeat steps 9 thorough 11
9. Read S
10. Identify the keywords in S
11. If S has no keywords
    Then
        Discard S
    Else
    i. Annotate S with c
    ii. Evaluate the entropy of S using (16)
12. Group sentences into granules with respect to the subcontext annotation
13. Arrange the sentences in the granules in descending order of their entropy
14. Stop

Note:

(a) The algorithm is part of the preprocessing function within the encoder (Fig. 1).
(b) As the steps of pos-annotation, pronoun–noun resolution and stop-word removal have been included in this algorithm, there is no need to repeat the same for the extraction of the Z-valuation parameters.



(a) The keywords w_1, w_2 and w_3 in a paragraph belong to the contexts (c_i) as shown and it would not be incorrect to assume that the paragraph pertains to the subcontext c_3.



(b) The subcontext of discourse of the paragraph containing sentences S_1 and S_2 is c_1.

Fig. 2. Pictorial depiction of the process of identification of the subcontext of discourse.

(c) The algorithm reasons that a sentence without a single keyword is considered irrelevant. However, the pronoun resolution helps relate sentences within a granule.

E.g. after pronoun resolution, the following text sample,

"The table needs to be rearranged. It looks rather untidy" (State I) is transformed to

"The table needs to be rearranged. The table looks rather untidy" (State II), where 'table' is a keyword in some context.

Thus, while in State I, where the first sentence is considered relevant and the second is not, both the sentences in State II are relevant.

(d) The removal of stop-words is done after the resolution of the pronouns so as to ease the identification of the inter-relevant sentences. For a list of common stop words in English, one may refer to [47].
(e) The complexity of the above algorithm depends on the following factors:
   i. The algorithm followed for pos-annotation.
   ii. The algorithm followed for the pronoun–noun resolution.
   iii. The algorithm used to remove the stop words.
   iv. The algorithm used to extract and associate keywords with subcontexts.
   v. The data structure of the text-corpus dictates the complexity of the storage and retrieval of its contents.
(f) The algorithm not only contributes to the identification of important sentences in a text sample for removal of the irrelevant, but also to the initial granulation of concepts within the text.
(g) The algorithm begins with the identification of the inter-relevance of sentences in a paragraph (as in step 6), to then proceed with the identification of the intra-relevance (step 10).
(h) The algorithm hints at the incorporation of commonsense computing perspectives during the association of the keywords to the subcontexts (step 6).
(i) The algorithm is adaptive (by virtue of step 7).
(j) Fig. 2 depicts pictorially the logic underlying the identification of the subcontexts in a given text sample.

At this juncture, one might argue that the prior probabilities of the subcontexts might dictate the course of comprehension; this however, is negated by the consideration of the inter-relevance association between the keywords in a paragraph.

5. Experiment

5.1. Experimental set-up

Considering the problem of 'text-identification' described in [4], the logic described in Section 4.2, was applied on a set of text-excerpts (synopses) from different Agatha Christie mysteries.

After being trained by 30 works of fiction (genre: mystery), the machine (M) possessed a vocabulary of 165 basic keywords [4] (not considering their different morphological forms) lying under two broad subcontexts: 'murder mysteries' and 'others'.

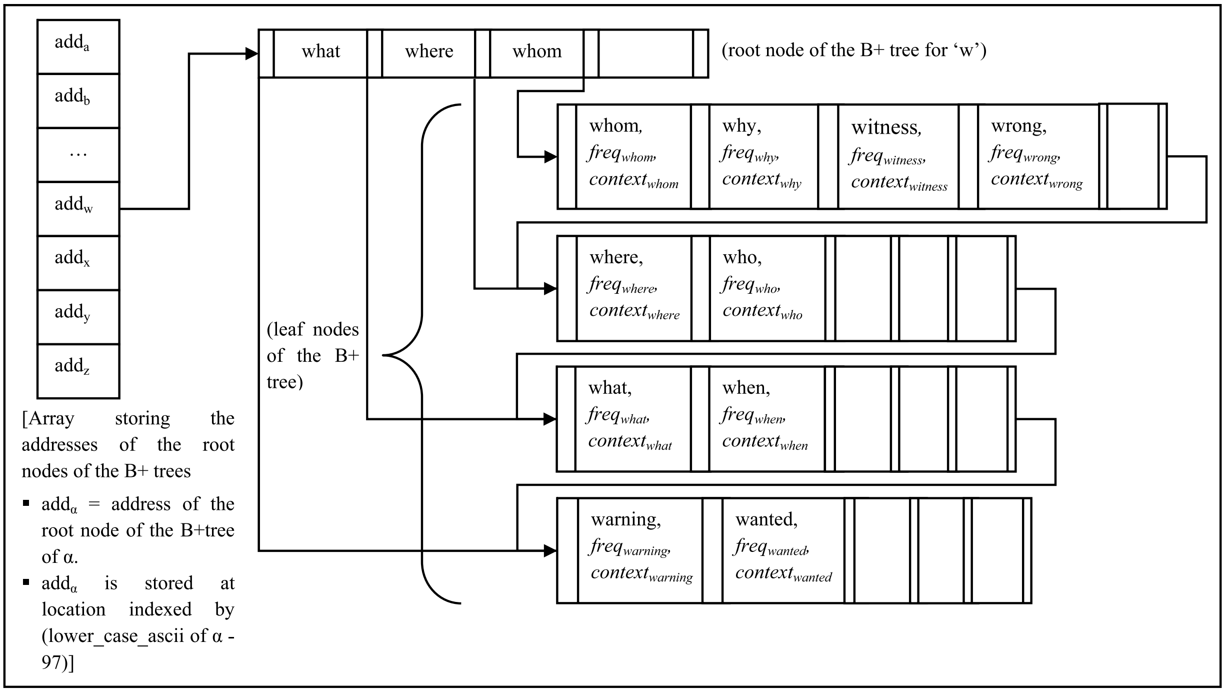


Fig. 3. The corpus storage structure.

The keywords comprise nouns, noun-phrases, adjectives and verbs that occur significantly across the training texts. The list also includes the names of the detectives and the police personnel.

The entropy-based relevance-ranking of the sentences, as yielded by our algorithm, was validated against the perceptions of 15 human subjects in the age-group 25–30 with varying reading habits and across a number of professions. Since our algorithm basically seeks to formalize an inherently intuitive process, the validation helps match the performance of our algorithm against human beings in general.

The subjects were asked to read the given sample, extract the words they thought were keywords and order the sentences from the most important to the least, also stating their reasons for doing so. A study of the reasons revealed that the subjects actually chose the sentences that, they deemed, provided sufficient information regarding the content of the entire text.

The experiment used standard schemes for pos tagging [43], pronoun resolution [44], and stop-word removal, while keyword association to subcontexts was done using WordNet [45,46] and ConceptNet [15]. The data structure to maintain the word-corpus is where our contribution lies and thus the article includes a note on it.

A note on the data structure used for the corpus

In a B+ tree of order b with h levels [root is at level 1], the maximum number of nodes ($n_{\max}$) that can be stored is given as,

$$n_{\max} = b^h - b^{h-1}. \tag{17}$$

Thus, assuming there would be a maximum of 100 words, per alphabet, the corpus is maintained as a forest of 5-order B+ trees – a B+ tree per alphabet – for the following reasons:

(a) With increasing data, a B+ tree does not become unwieldy like an indexed-sequential file;
(b) In the early stages of execution, when the machine requires learning new words, the corpus experiences a greater number of insertions than search operations. Insertion in a B+ tree employs small local reorganizations, unlike a B tree;
(c) A B+ tree has a smaller number of levels compared to a binary search tree. The number of key-comparisons is equal to the number of levels;
(d) All the leaves are at the same level. Thus, the time to locate a leaf-node is nearly the same, irrespective of the search value; the leaf-nodes hold the keywords and the associated information; The number of nodes that need to be searched to locate the appropriate leaf-node $= O(\log_b n)$;
(e) We do not consider the possibility of representing the notion of 'forgetfulness' by simulating the gradual deletion of a word from a tree.

The addresses of the root nodes of the B+ trees are saved in an array — in alphabetic order. Each of the nodes in the tree is a linked list, the header of which holds the number of entries in the node. The structure of the corpus is shown in Fig. 3.

Complexity of the basic operations with respect to the data structure used:

(a) Insertion includes the following elements –
  i Time to look up root-node address: $O(1)$
  ii Time to locate leaf node: $O(\log_b n * \log_2 b) = O(\log_2 n)$
    [Time to search within a node $= O(\log_2 b)$]
  iii Time to reorganize tree:
    [*Best case:* No tree reorganization required
    Average case: No new levels are created, therefore,
    Time to count number of entries in leaf node $= O(1)$
    Time to locate median of node $= O\left(\frac{b}{2}\right)$
    Time to update parent node $= O(1)$ (new key entered at the end of list)
    Time to split leaf node into two new leaf nodes and pointer adjustments $= O(c)$
    Worst case: A new level is created, i.e. parent nodes all the way to the root are required to be split and accordingly updated. Thus, all the steps of the average case need to be executed $O(\log_b n)$ times]
  iv. Time to insert node into leaf node: $O(\log_2 b)$.
(b) Searching is composed of the following sub-operations –
  i. Time to look up root-node address: $O(1)$
  ii. Time to locate leaf node: $O(\log_2 n)$.

5.2. Results

Table 1 tabulates the text excerpts considered, results of our algorithm and the validation process. The keywords have been italicized; the algorithm outputs the sentences of the text in decreasing order of importance and the validation depicts the percentage out of a total of 15 human subjects who agreed on the algorithm output.

As is evident, the algorithm shows an average 71.65% match with the opinions of the subjects. We thus infer that, our algorithm, though not perfect (e.g. the word '*wrong*' is incorrectly considered a keyword in sample (4)), is certainly capable of identifying the sentences of importance, i.e. those that are relevant to the comprehension of the excerpt.

6. Conclusion

The concepts described in this paper are a product of our research on 'Computing for Cognition' that seeks to integrate paradigms like computing with words, natural language processing, affective computing, and commonsense computing to design a framework for cognitive architectures. What we aspire to achieve is a machine that possesses commonsense, is able to think, learn and apply knowledge to solve very specific (e.g. learning to cross the road) problems, interact with an autistic child or assist Alzheimer's patients. The Z-number is capable of integrating the CWW, NLP and affective computing paradigms, while the proposed algorithm tries to encompass elements of commonsense computing.

This paper is an elucidation on our investigations on the modeling of the neural 'relevance-cognition' process—instigated by our algorithm for the Z-number approach to CWW. We present here a subjectivist corpus-based logic, that draws from the principles of Bayes' theorem and Shannon's entropy theorem, to granulate concepts and rank sentences in a text sample in the order of the importance of the information they convey; all within the purview of a context. The proposed methodology, though lies in conjunction with our investigations on the Z-number, could influence the research on text summarization, disambiguation of possible multi-contextual interpretation of text and concept graph extraction.

The results obtained, in the course of the experiments, are quite encouraging, considering an average of 71.65% match with human perceptions. These show that the entropy of a sentence does indeed possess the ability to model the subtle perceptions that affect the comprehension of a text. What remains to be seen is if the concept applies to any given context – whether it be clinical diagnosis, bank transactions or risk analysis – areas that involve intensive transactions with human beings. The directions in which our methodology might be improved are:

(a) The machine should be able to extract the sentences that are most important while discarding the irrelevant. This supposedly calls for the concept of an entropy-threshold value.
(b) The machine's capability of identifying subcontexts under a particular context depends on the system-designer's knowledge of the same. The algorithm could be enhanced to be able to identify and handle 'new' subcontexts too.
(c) The algorithm assumes all sentences to be syntactically and semantically correct, and thus requires being sufficiently robust so as to be able to handle otherwise.
(d) The algorithm considers each sentence to be an individual entity within a granule. The algorithm should also be able to also extract the combined information from all related sentences [48].
(e) Incorporation of measures of truth-values of the information extracted from the sentences requires attention [49]
(f) Experimentation over the data structures storing the word-corpus, inspired by cache-based retrieval schemes would surely be appreciated, considering the continually growing vocabulary.
(g) Adapting the algorithm to process subtle language usage as in the case of processing the question, "How many survivors of a plane crash were buried?'' In its current form, the machine being greater acquainted with the concept of 'buried $\rightarrow$ death' overlooks the importance of the concept of 'survivor $\rightarrow$ alive' in the sentence.

Seeing to the incorporation of these elements summarizes the next phase of our work.

Table 1
Results showing original text samples, order of the sentences on the basis of the entropy and the percentage match with the perceptions of human subjects. (The keywords are italicized.)

Original text sample	Sentence order on the basis of the entropy	% match with human subjects
(1) Hercule *Poirot* does not need all his *detective skills* to *realize* something is *troubling* his secretary, Miss Lemon—she has made three *mistakes* in a simple letter. It seems an outbreak of *Kleptomania* at the student hostel in which her sister works is distracting his usually efficient assistant. *Deciding* that *desperate* times call for *desperate measures*, the great *detective decides to investigate*. *Unknown* to *Poirot*, however, *desperation* is a *motive* he shares with a *killer*. (Hickory Dickory Dock)	(Subcontext identified: murder mystery) (a) It seems an outbreak of *Kleptomania* at the student hostel in which her sister works is distracting his usually efficient assistant. (b) *Unknown* to *Poirot*, however, *desperation* is a *motive* he shares with a *killer*. (c) Hercule *Poirot* does not need all his *detective skills* to *realize* something is *troubling* his secretary, Miss Lemon—she has made three *mistakes* in a simple letter. (d) *Deciding* that *desperate* times call for *desperate measures*, the great *detective decides to investigate*.	60% (The human subjects who agree to the ordering are of the opinion that sentences (a) and (b) lead to the conclusion that the story deals with 'thieves' and 'murderers'. Sentences (c) and (d) are irrelevant)
(2) For an instant the two trains ran together, side by side. In that frozen moment, Elspeth was *sure* she *witnessed* a *murder*. *What* else could it have been? As she stared helplessly out of her carriage window, a man had *remorselessly strangled* a woman. The *body* crumpled. Then the other train drew away. But *who* apart from Jane *Marple*, would take her story seriously? After all, there are no *suspects*, no other *witnesses* and no *corpse*. (4.50 from Paddington)	(Subcontext identified: murder mystery) (a) As she stared helplessly out of her carriage window, a man had *remorselessly strangled* a woman. (b) In that frozen moment, Elspeth was *sure* she *witnessed* a *murder*. (c) After all, there are no *suspects*, no other *witnesses* and no *corpse*. (d) The *body* crumpled. (e) *What* else could it have been? (f) But *who* apart from Jane *Marple*, would take her story seriously?	73.3% (The human subjects who agree to the ordering are of the opinion that sentences (a), (b) and (c) lead to the conclusion that the story deals with a 'murder'. The other sentences are irrelevant)
(3) Lymstock is a town with more than its share of *secrets*—a town where even a sudden outbreak of *anonymous hate-mail* causes only a minor stir. But all that changes when one of its recipients, Mrs. Symmington, *commits suicide*. Her final note said, "I cannot go on". Only Miss *Marple questions* the *coroner's verdict* of *suicide*. *Was* this the work of a *poison*-pen? Or of a *poisoner*? (The Moving Finger)	(Subcontext identified: murder mystery) (a) But all that changes when one of its recipients, Mrs. Symmington, *commits suicide*. (b) *Was* this the work of a *poison*-pen? (c) Or of a *poisoner*? (d) Lymstock is a town with more than its share of *secrets*—a town where even a sudden outbreak of *anonymous hate-mail* causes only a minor stir. (e) Only Miss *Marple questions* the *coroner's verdict* of *suicide*.	73.3% (The human subjects who agree to the ordering are of the opinion that sentences (a), (b) and (c) lead to the conclusion that the story deals with a 'possible murder'. The other sentences are irrelevant)
(4) When Miss *Marple* comes up from the country for a holiday in London, she finds what she is looking for at Bertram's Hotel: traditional décor, impeccable service—and an unmistakable atmosphere of *danger* behind the highly polished veneer. Yet, not even Miss *Marple* can foresee the *violent* chain of events set in motion when an eccentric guest makes his way to the airport on the *wrong* day. (At Bertram's Hotel)	(Subcontext identified: other) (a) Yet, not even Miss *Marple* can foresee the *violent* chain of events set in motion when an eccentric guest makes his way to the airport on the *wrong* day. (a) When Miss *Marple* comes up from the country for a holiday in London, she finds what she is looking for at Bertram's Hotel: traditional décor, impeccable service—and an unmistakable atmosphere of *danger* behind the highly polished veneer.	80% (The human subjects who agree to the ordering are of the opinion that sentence (a) creates a stronger sense of mystery than sentence (b))

Acknowledgments

This project is being carried out under the guidance of Professor Sankar K. Pal who is a J.C. Bose Fellow of the Government of India.

References

[1] J.B. Tenenbaum, C. Kemp, T.L. Griffiths, N.D. Goodman, How to grow a mind: statistics, structure, and abstraction, Science 331 (11) (2011) 1279–1285.
[2] N. Fang, X. Luo, W. Xu, Measuring textual context based on cognitive principles, International Journal of Software Science and Computational Intelligence 1 (4) (2009) 61–89.

[3] M.C. Frank, N.D. Goodman, J.B. Tenenbaum, Using speakers' referential intentions to model early cross-situational word learning, Psychological Science 20 (2009) 578–585.
[4] S.K. Pal, R. Banerjee, S. Dutta, S.S. Sarma, An insight into the Z-number approach to CWW, Fundamenta Informaticae (2012) (special issue: Cognitive informatics and computational intelligence) (in press).
[5] T.F. Jaeger, Redundancy and reduction: speakers manage syntactic information density, Cognitive Psychology 61 (2010) 23–62.
[6] D. Genzel, E. Charniak, Entropy rate constancy in text, in Proc. of the 40th Annual Meeting of the Association for Computational Linguistics, ACL, Philadelphia, 2002, pp. 199–206.
[7] D. Genzel, E. Charniak, Variation of entropy and parse trees of sentences as a function of the sentence number, in Proc. 2003 Conference on Empirical Methods in Natural Language Processing, EMNLP, Stroudsburg, PA, USA. Association for Computational Linguistics, 2003, pp. 65–72.
[8] F. Keller, The entropy rate principle as a predictor of processing effort: an evaluation against eye-tracking data, in Proc. of the Conference on Empirical Methods in Natural Language Processing, 2004, pp. 317–324.
[9] T. Qian, T.F. Jaeger, Close = relevant? The role of context in efficient language production, in Proc. of the 2010 Workshop on Cognitive Modeling and Computational Linguistics, Uppsala, Sweden. Association for Computational Linguistics, 2010, pp. 45–53.
[10] T. Qian, T.F. Jaeger, Topic shift in efficient discourse production, in The 33rd Annual Meeting of the Cognitive Science Society, CogSci11, Boston, MA, 2011, pp. 3313–3318.
[11] P. Gupta, V.S. Pendluri, I. Vats, Summarizing text by ranking text units according to shallow linguistic features, in Proc. 13th International Conference on Advanced Communication Technology, ICACT, 2011, pp. 1620–1625.
[12] P.F. Brown, P.V. deSouza, R.L. Mercer, V.J.D. Pietra, J.C. Lai, Class-based n-gram models of natural language, Computational Linguistics 18 (4) (1992) 467–479.
[13] L.A. Zadeh, Fuzzy logic = computing with words, IEEE Transactions on Fuzzy Systems 4 (2) (1996) 103–111.
[14] L.A. Zadeh, Intelligent systems revolution: is it real? http://www-bisc.cs.berkeley.edu/zadeh/papers, 1998.
[15] C. Havasi, R. Speer, J. Alonso, ConceptNet 3: A flexible, multilingual semantic network for common sense knowledge, in: Proc. Recent Advances in Natural Languge Processing, 2007.
[16] L.A. Zadeh, Outline of a new approach to the analysis of complex systems and decision processes, IEEE Transactions on Systems, Man and Cybernetics 3 (1) (1973) 28–44.
[17] L.A. Zadeh, Test-score semantics for natural language and meaning representation via pruf, in: B. Reiger (Ed.), Empirical Semantics, Dr. Broackmeyer University Press, Germany, 1982, pp. 281–349.
[18] L.A. Zadeh, Precisiated natural language (PNL), AI Magazine 25 (3) (1994) 74–91.
[19] L.A. Zadeh, A new direction in AI: toward a computational theory of perceptions, AI Magazine 22 (1) (2001) 73–84.
[20] M. Delgado, O. Duarte, I. Requena, An arithmetic approach for the computing with words paradigm, International Journal of Intelligent Systems 21 (2006) 121–142.
[21] J. Lawry, A methodology for computing with words, International Journal of Approximate Reasoning 28 (2001) 55–89.
[22] G. Trivino, A. Vander-Heide, An experiment on the description of sequences of fuzzy perceptions, in: Proc. 8th International Conference on Hybrid Intelligent Systems, Barcelona, Spain, IEEE Press, NJ, 2008, pp. 228–233.
[23] A. Vander-Heide, G. Trivino, Simulating emotional personality in human computer interface, in: Proc. IEEE International Conference on Fuzzy Systems, Barcelona, Spain, IEEE Press, NJ, 2010, pp. 1–7.
[24] J.M. Mendel, D. Wu, Perceptual Computing — Aiding People in Making Subjective Judgements, IEEE Press, NJ, 2010.
[25] M. Reformat, C. Ly, Ontological approach to development of computing with words based systems, International Journal of Approximate Reasoning 50 (2009) 72–91.
[26] R.R. Yager, On the retranslation process in Zadeh's paradigm of computing with words, IEEE Transactions on Systems, Man and Cybernetics, Part B (Cybernetics) 34 (2) (2004) 1184–1195.
[27] J. Kacprzyk, S. Zadrozny, Computing with words is an implementable paradigm: Fuzzy queries, linguistic data summaries and natural language generation, IEEE Transactions on Fuzzy Systems 18 (2010) 461–472.
[28] L.A. Zadeh, A note on Z-numbers, Information Sciences 181 (2011) 2923–2932.
[29] C.E. Shannon, A mathematical theory of communication, Bell system Technical Journal 27 (3) (1948) 379–423.
[30] K. Bach, Context dependence (such as it is). [Online]. Available: http://userwww.sfsu.edu/~kbach/.
[31] E.J. Gibson, Perceptual learning and the theory of word perception, Cognitive Psychology 2 (1971) 351–368.
[32] J. Erik-Mai, Classification in context: relativity, reality, and representation, Knowledge Organization 31 (1) (2004) 39–48.
[33] F.J. Pelletier, Context dependence and compositionality, Mind and Language 18 (2) (2003) 148–161.
[34] P. Lasersohn, Context dependence, disagreement, and predicates of personal taste, Linguistics and Philosophy 28 (2005) 643–686.
[35] L.W. Barsalou, Context-independent and context-dependent information in concepts, Memory and Cognition 10 (1) (1982) 82–93.
[36] K. Bach, Semantic slack: what is said and more, in: S.L. Tsohatzidis (Ed.), Foundations of Speech Act Theory, 1994.
[37] F. Recanati, What is said, Synthese 128 (2001) 75–91.
[38] K. Bach, Knowledge in and out of context. [Online]. Available: http://userwww.sfsu.edu/~kbach/.
[39] N. Fang, X. Luo, W. Xu, Evaluating text understanding based on context, in: Y. Wang, D. Zhang, J.C. Latombe, W. Kinsner (Eds.), Proc. 7th IEEE Int. Conf. on Cognitive Informatics, IEEE Press, NJ, 1998, pp. 195–204.
[40] [Online]. Available: http://www.trinity.edu/cbrown/bayesweb/.
[41] [Online]. Available: (Stanford Encyclopaedia of Philosophy) http://plato.stanford.edu/entries/bayes-theorem/.
[42] S.M. Ross, Introduction to Probability Models, ninth ed., Academic Press, London, UK, 2007.
[43] E. Charniak, Statistical techniques for natural language parsing, AI Magazine 18 (4) (1997) 33–44.
[44] S. Lappin, H.J. Leass, An algorithm for pronomial resolution, Computational Linguistics 20 (4) (1994) 535–561.
[45] G.A. Miller, WordNet: a lexical database for english, Comunications of the ACM 38 (11) (1995) 39–41.
[46] C. Fellbaum (Ed.), WordNet: An Electronic Lexical Database, MIT Press, Cambridge, MA, 1998.
[47] [Online]. Available: http://norm.al/2009/04/14/list-of-english-stop-words/.
[48] R. Carnap, Y. Bar-Hillel, An outline of a theory of semantic information, Research Laboratory of Electronics, Massachusetts Institute of Technology, Tech. Rep. 247, 1952.
[49] L. Floridi, Is semantic information meaningful data? Philosophy and Phenomenological Research 70 (2) (2005) 351–370. [Online]. Available: http://philsci-archive.pitt.edu/2536/1/iimd.pdf.