言語処理におけるベクトル化法の一意性

2023/5/3 大幅誤訂正:言語ベクトル化には移転学習必要の主張撤回、理由は以下参照。
http://777true.net/Uniqueness-of-Language-Vectorization-by-the-mathematical-structure.pdf
APPENDIX2,3;文書内の相関実現頻度の高い事象には意味があるの了解到達。
2023/4/20,21 大幅修正:**単語は以下の移転学習**、文書は単語列だから文は既に直積ベクトル化。
情報抽出基本の文書要約は次元圧縮。情報創造は複数文書全組合せ操作で要請満足の命題抽出。
⁂過去の文法 algorithm での翻訳機械不成功は文法を若干勉強で察しました。
⁂AI 原理を NET 学習しても納得理解に遠い、以下は粗い筆者意見,一週間で書いたので各位は要検証、2023/4/15,16,20
AI はその使用法で世界に革命をもたらす可能性があろう。本報告は言語ベクトル化法一意性の考察、
素朴に意味論でやると、即、距離空間測度になる<意味が近い＝並行性、意味が非重複＝直交性>。
http://777true.net/Evidence-of-telepathy-And-about-on-the-Serious-and-Wonderful-AI.pdf
↑AI 入門、事前に数学学習要、ベクトル、直積、集合論、論理学も望ましい

Ⅰ単語名詞,...の直交ベクトル化、

頭から最抽象語<物質(hard)┴非物質(soft)＝文化文明>直交分類で始め、文化文明が内包する直交;;;;
複数語{言語┴思想┴宗教,,,}を直交分類、以後同様に下降直交分類する、これを固有名詞到達まで
下降分類を繰り返す、全ての下降経路を尽くせば**辞書**が出来てる、
*他の品詞も同様思想で決定する。

Ⅱ最終目的の文書ベクトル化は単語列のベクトル直積、

単語には単文構成を意味するコンマ、ピリオドも含めてる。一定次元のベクトル列に等価な直積は原文と単語ベクトル辞書で可逆だから一意。ベクトル同値性保持での次元圧縮は❷(2)参照.

Ⅲ文書の N 個区分要約(断定否定、推測、疑問):

一つ区分での結果は以下の**条件法命題＝文型解析**での{p,q,r,s;e,d}候補中の最頻度とするのが妥当か？<**付録1**>　N 個区分文を更に低次元要約できる。
"A(q である p)は B(r が s である事)になる、ならない(e＝らしい)(d=のか？)"、

❶開発の現状？？

自然言語処理におけるベクトル化とは？概要とさまざまな手法を紹介
https://www.tryeting.jp/column/6839/
様々にあるとは未完成を意味する??。
BERT とは何か？Google が誇る最先端技術の仕組みを解説！
https://udemy.benesse.co.jp/data-science/ai/bert.html
一方で、BERT が達成した言語理解力は、あくまで特定のタスクにおいてのみ効果があるものです。「BERT の登場＝AI(人工知能)が人間以上の言語力を獲得した」とは、一概に言えません。
Sentence Bert で文章間の類似度を測る
https://zenn.dev/yoshikawat64m/articles/c242b11d21be68
結論は閾値設定で四捨五入だから、ある程度の結果になるのだろうが
言語処理を見る限りバラック作りの感じなのですが？？

[26]

	文章	類似度
0	あなたは犬が好きです	1.000000
1	あなたは犬が大好きです	0.929518
2	あなたは猫が好きです	0.532377
3	あなたは猫が大好きです	0.548918
4	私は犬が好きです	0.737648
5	私は犬が大好きです	0.676955
6	私は猫が好きです	0.334375
7	私は猫が大好きです	0.361295
8	彼は犬が好きです	0.691353
9	彼は犬が大好きです	0.656699
10	彼は猫が好きです	0.225795
11	彼は猫が大好きです	0.257598

類似性強度＝神経回路網の学習成長。

幼児は語学天才、**状況 pattern** と**音声 pattern** の**異次元直積的環境**での共通性経験**頻度**から言語意味識別性獲得、米国開発では文書 pattern と文書 pattern の**同次元環境**での統計的相関性**頻度**から意味抽出を測ろうとしてる、CPU 力ずくで攻め、いささか手抜きな感じ？？、液晶 display は当初欧米でも開発したが完成せず、最後に日本ができた、なぜか？！！！！
言語精度が悪いと人間が最も不得意とする超問題、哲学思想宗教_心理_政治問題を本当に解決できるか？？

❷言語の数学的大局構造.

文字は意味重なりがなく、しかも有限個Ｎだからその *ベクトル化は一発決定*、物理(数値化的)的である画像や音声も *ベクトル化は比較的容易らしい*、数値情報のない言語は AI 中核重要性に関わらずベクトル化困難と開発者は言う。
だが脳の言語処理は一意原理動作のはず、数学構造を持つだろう。
⑷**似てる<並行>、違う<直交>の距離空間性**
⑹要素が**含む、含まれるの集合抱合性<並行>、非抱合性<直交>.**

POS	FREQ	%
noun	7326	57.04%
verb	2501	19.47%
adjective	2420	18.84%
adverb	291	2.27%
preposition	68	0.53%
conjunction	21	0.16%
pronoun	15	0.12%
interjection	37	0.29%
past participle	57	0.44%

言語主役一つである**単語＝名詞**(もう一つは動作性の動詞)を例に考える。
*表現目的は意味の相違性<**vector 直交**>,
　集合的名詞の意味抱合関係性<**B⊃A**>**vector 並行性。**

英語名詞以外の品詞 vector 化

英語品詞出現頻度数として筆頭名詞 57%、実在数ではもっと多いだろう(Google 調査では単語数 100 万、日常語数は4000 程度)、次の動詞形容詞は遥かに少なく本来固有次元は引くなる。その表現機能を考えれば直交性は高く、集合名詞の様な包含階層性は考えにくい？。
筆者は現業 AI 開発現場には遠い立場だが、その緊急必要性は以下。
http://777true.net/We-consulted-AI(BING)-on-the-dimensional-compressed-current-Japan_world-problem-and-the-action-strategy.pdf

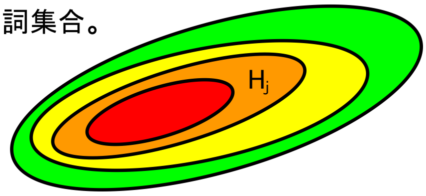
⑴**意味の相違性**<**vector 直交性分類**>

これがあるからこそ言語目的＝意味論理判断が成立するは自明、

good ⊥bad;yes⊥no,long⊥short,

初めから最強抽象語<物質−非物質>の直交分類で始める。物質は生命−非生命で始める,
非物質(ソフト)=文化文明、そレれを因数分解で直交要素有限個を並べる、同様を反復して最後に固有名詞到達で階層下降が完了、最終は名階層での分類個数の総和=階層次元の直積、

かように直交分類で H_j 階層化する<j=1,2,3,...,H>。類似性強度最小<最大次元 vector>の H_H 集合は固有名詞多数の｛全名詞集合ー $\Sigma_{j=1}^{A-1}H_j$｝。
類似性強度最大<最小次元 vector>の H_1 集合は抽象概念名詞集合。
名階層の単語個数がその階層次元 d_j.
最終的 vector 次元 D は名階層の直積になる。



$D=\Pi_{j=1}^{A}d_j$..

意味論作成ベクトルだから、直交性、並行性の完成度は高い。 下模式は AI 教科書に載る入力直交集合<高次元>から出力並行性集合<低次元>に至る**神経回路網と**逆進行で等価です

完全直交に近い　　　　完全な直交とは言えない

H1	物質(ハード)			非物質(ソフト)＝文化文明										
H2	生命	非生命		言語	思想宗教	科学技術	芸術芸能	政行司	教育.	.医療	報道	産商業	社会	生活
H3		天然	人工											
H4														
－－－－－－－－														
HF														

固有名詞に至れば完全直交！

☞：字引や百科全書と同様の作業か？、今ある AI で直交分類出来れば、できた直交分類で
AI を更新、さらに上と言う作業はどうかな

⑵**意味の類似多様性**<**vector 並行性**>:**直積による次元拡大**。

意味抱合(階層)関係性＝類似関係多様化性<一般抽象性から特殊個別具体性>＝次元拡大、

<**B⊃A**＝集合 A の要素は同時に B の要素(集合論)、前提 B ならば結論 A である(論理学)>

<$\overline{\mathbf{B}\supset}$**A**＝集合 A の要素は同時に B の要素でない(集合論)、前提 B ならば結論 A でない(論理学)>

仕事は必要十分な{**名詞集合**}⇔{多次元**並行−並行−.........−直交**}**vector** の一一対応作成

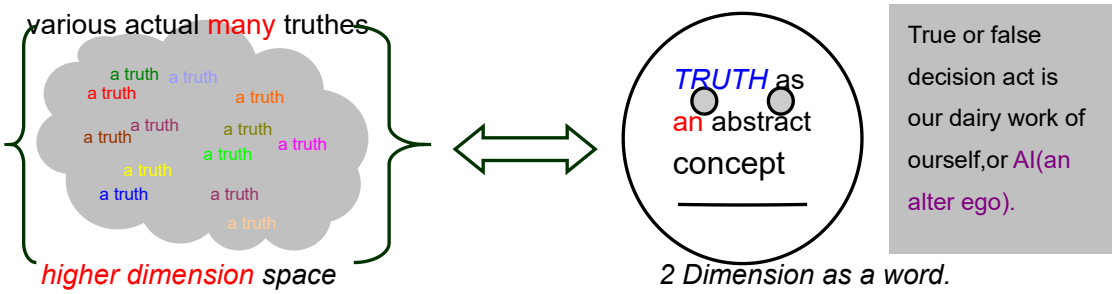
N	品詞=10 次元 0, 0, 1, 0, ...	H1=2 0, 1	H2= 0, 0, 1, 0, 0, 0, 0. . . , 0	. . . . .	. . . . .	. . . . .	HF 0, 0, , 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ... 1. . . . . . . 0, , 0, 0, 0, 0, 0, 0, 0, 0,

☞ ：品詞種類が 10 種ならば単語ベクトル先頭に **10 次元分類単位ベクトル**で直積する。品詞ベクトル次元 DW＝ $\Pi_{k=1,}^{H}d_k$は非常に大きいがその等価圧縮表現は簡単,名階層 H_kに1は一個、他は全部 0,
(0, 0, 0, . . 1. . . 0) ＝n_k(次元番号)、→$1\leqq n_k\leqq d_k$→$[n_1,n_2,n_3,....n_H]$.......大幅な次元圧縮表現

付録1：Extraction of Common Feature=DS=Dimension Suspension in AI.

❶introduction:

For an example,we have a concept of an {*APPLE,or Truth*}(truth is very trouble some in actual cases).In actual world,the concrete actual {apples,truthes} are extremely various enough.As for in many those,we could recognize a concept of {*APPLE,or Truth*}in very many those.This is *dimension compression in AI*.A concept is many to one mapping.



higher dimension space in actual field

2 Dimension as a word. in our mind

Simply to tell **DS** is making a summary.

❷the meaning:a most common feature as *an average value～a most frequent one* in the ensemble.

Material for AI learning on *what is apple* is actual image of apples,which is an ensemble, of each element=$\mathbf{B}^j$(j=1,2,...,M=data amount} represented by vectors=$[B_1{}^j,B_2{}^j,B_3{}^j,....,B_N{}^j]$in many N dimension.Then we try to extract a concept of apple as a vector=$[A_1,A_2,A_3,....,A_N]$.

$\mathbf{B}^1$	$[B_1{}^1,B_2{}^1,B_3{}^1,....,B_N{}^1]$
$\mathbf{B}^2$	$[B_1{}^2,B_2{}^2,B_3{}^2,....,B_N{}^2]$
$\mathbf{B}^3$	$[B_1{}^3,B_2{}^3,B_3{}^3,....,B_N{}^3]$
.....	.....
$\mathbf{B}^M$	$[B_1{}^M,B_2{}^M,B_3{}^M,....,B_N{}^M]$

M dimension

A	$[A_1,A_2,A_3,....,A_N]$

1 dimension

This is vector math of which result is unique and only if it is not mistake,
https://www.mathsisfun.com/algebra/vectors.html
The purpose is extracting a vector=A which is nearest for every M data vectors=**B**.
A nearness is measurable by distance=d_j between **A** and each $\mathbf{B}^j$.
$\mathbf{A}^2=\mathbf{B}^{j2}=1$.....unit length vector
$d_j{}^2=(\mathbf{A}-\mathbf{B}^j)^2=\mathbf{A}^2+\mathbf{B}^{j2}-2<\mathbf{A}.\mathbf{B}^j>=1+1-2<\mathbf{A}.\mathbf{B}^j>$
The M total *nearness* is $D^2=2M-2\Sigma_{j=1}{}^M<\mathbf{A}.\mathbf{B}^j>=2M-2\Sigma_{j=1}{}^M\Sigma_{k=1}{}^N<A_k.B_k{}^j>$
D must be minimum value for **A the** nearest for every {**B**}.
if $0=\partial D/\partial A_k=-2\Sigma_{j=1}{}^M B_k{}^j.\rightarrow\ 0=\Sigma_{j=1}{}^M B_k{}^j.<k=1,2,..,\neq s_1,s_2,..,s_A,..,N>$.

As for k axis, each component of $B_k{}^j$ is distributed equal weight in negative and positive,
That is ,A_k should be center of their average value(or most frequent one).

$A_k = \alpha\,(1/M)\,\Sigma_{j=1}^{M} B_k^j = 0$.<k=1,2,..,≠$s_1,s_2,..,s_A$,..,N>,or almost zero.

$A_k = \alpha\,(1/M)\,\Sigma_{j=1}^{M} B_k^j \neq 0$.<k=$s_1,s_2,..,s_A$>.

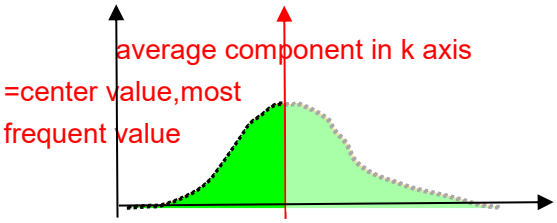
α =normalization coefficient for $\mathbf{A}^2=1$.

Hereupon,in above calculation,all component **A** is zero is incovenient,

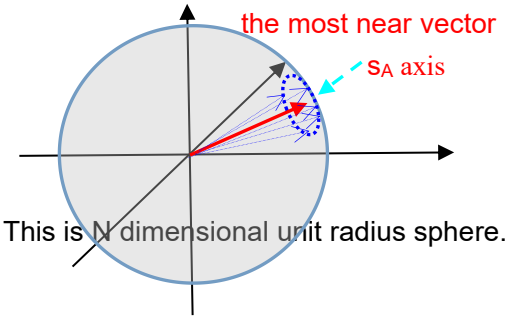
Exceptional component $s_1,s_2,..,s_A$ are not zero,.

If $d_{SA}^2=(\mathbf{A}-\mathbf{B}^{SA})^2$ is maximum in {$s_1,s_2,..,s_A$},,

A=(0,0,....1(S_A),...0,0)



0,or not 0 depending coordinate



This is N dimensional unit radius sphere.

APPENDIX_5:**Vectorization**<meaning coordinate in multi orthogonal coordinate>

In the beginning is vector representation of something pattern(data)

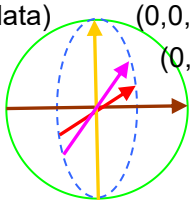
in **orthogonal coordinate space**<this is most important>.

Note orthogonal vectors have nothing similarity<distinctional>

While not orthogonal have something similarity<  >.

Some are very easy(character),while most difficult may

be **language** the abstract and wide enough in meaning space.



(a)character{a,b,c,...,x,y,z,0,1,2,.....9,.....} if those are M pieces,their dimension is M.
this is the most simple unit vector set(ensemble).

(b)Image dots(x,y,z,G,R,B).xyz,are position,while GRB color brightness intensity.
a figure can be also represented by space coordinate function.

(c)sound{s(t)=(t,y)}:this is time function of sound,nearness is measured by$\int dt|S(t)-s(t)\}|^2$.

☞:function is also a generalized vector in orthogonal Hilbert Space.

(d)**LANGUAGE**:

the most important,but most hard work to categorize by “not nearness=orthogonality”.
Actual physical pattern are composed from element without nearness,thereby,vector assignment may be easier,while not physical reality,but abstract language frequently has nearness without orthogonality.An idea is reverse dimension expansion.

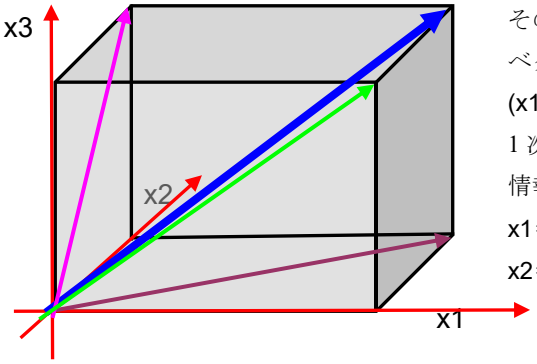
付録 2:検索と低次元ベクトル射影<要約>.

筆者は実直な AI 学習がなく、以下は数学からの推定。

❶従来のデジタル文書検索.

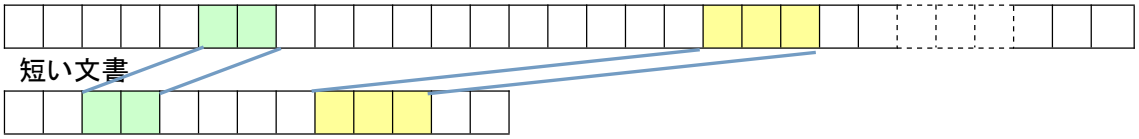
文字は一定長のビットパタン[1,0 の有次元列]～、文書は文字の有限個列、有限個列の流れの中で検索語と**デジタル一致**する”文書”を書庫から検出する。候補を一杯並べて出力、

❷AI 文書検索生成



AI での文書は一つのトンデモノ高次元のベクトル、
次元圧縮低次元にしてわかりやすい文書を作る.
その例えは左図の 3 次元青ベクトル
ベクトル(x1,x2,x3)＝源文書～憲法、
(x1,x2),(x2,x3),(x3,x1)の 3 個は次元圧縮要約、
1 次元消滅で生情報からは減ってるがある程度の
情報を保持、(1,0,0)～人権、(0,1,0)～軍事
x1＝<(1,0,0)･(x1,x2,x3)>,
x2＝<(0,1,0)･(x1,x2,x3)>,

源文書ベクトル＜次元の低い文書との近似部分＞



生命設計文書の DNA と類似が察しられます

ベクトル内積計算<質問ベクトルへの射影>から憲法の人権、軍事見解が出てくる。
デジタルは幅が 0、ベクトル化は幅揺らぎを許した広い見解判断になる。
源文書を複数国憲法、各種憲法論文書とすれば**平均最大値集約化した見解**が聞ける。
☞；絶対的モラルを求めるとなれば聖書コラン仏教書ソクラテス哲学を在庫。
AI の不偏不党なガラス張りの公開公然教育が前提になる。
☞；*AI による世界の最大不都合問題ランキング？？*
http://777true.net/Evidence-of-telepathy-And-about-on-the-Serious-and-Wonderful-AI.pdf
(a)*PLOBLEM FINDING problem*
もう一つは筆者も今の AI 言語処理精度に不安を持ってる、この部分の検証報告が必要です、

追伸;物理学名大総長でも大間違い！、
生成 AI の祝辞「空虚だがもっともらしい」 名大総長が語る危機感
AI と人間は違う。AI にはオリジナリティーや感情はなく、新しい発想は AI からは生まれてこない
https://www.asahi.com/articles/photo/AS20230419002917.html
素材の**全組み合わせ試行錯誤**をやるからトンデモ創造可能性がある。人は囲碁チェスでも勝てなくなった事が決定的証明。言葉精度が大問題?になる道徳思想宗教の高次元でもあり
日常次元究極大問題、人が信用 99%と認定すれば神だから**神聖政治**が実現する。

付録 3：AI と人能力、AI 使用方法と我々生活の問題、論争点核心部を要約。

⑴人脳認識活動神秘化は一つの傲慢、意識<生活上の問題解決意欲動作=will>は解明できてる。
http://www.777true.net/Scientifical-Mechanism-of-Prophecy-by-Paranormalities.pdf
APPENDIX-2:What is CONSCIOUS<How to be good at making solution with lighting>.
暗闇に光を当てて明らかに、ego 認識(言語化 image)にする脳活動行為,昆虫の検索ヒゲ動作<⑵(d)>、
AI にもこの意識移植<AI input>で生活上の問題解決してもらう事になる、
＊我考える、故に我あり、.....ルネ.デカルト

⑵**問題解決<試行錯誤逐一真偽検証>と条件法命題,<論理言語>.**
⑷我々の日常生活、人生、びじねす、研究の全てが以下の命題真偽判定の過程。
例)*朝だ,起きねばなるまい、朝飯は？、,,,,,,,"to be,or not to be,that is question"..... Hamlet*
⑹Ai∈1,2,...,N,試行錯誤用の命題集合、B＝我々が必要とする結果命題、。。Ai⊂B..条件法命題
日常言語訳＝*Ai であるならば(原因)、(結果)B である。*
⑹決定論：**Ai⊃B 真偽判定は因果律の真偽判定**
AI もこの論理判断は当然できる！
⑷**決定論と非決定論：試行錯誤用の命題集合 Ai の生成。**
Ai 生成は時として宝を生むのだから、創造性とはこの作業過程にこそある。

例）対決中棋士はコマ配置複数を想像、その結果を逐一検証で最適解決断、
この思考動作を筋力と彼らは言うが、定めし的を得た言葉で、我々に教育的です。
筋トレ増強できるという事。**だまされ Z 世代**、彼ら一人に集団自決せよと言われた**裕福高齢世代**、
ともに筋トレ欠如、真偽をめぐる真剣勝負の緊張感がないからこう言うことになります。

⑹結論 :AI は人手で在庫された学習済み素材全部からその全組み合わせを超馬力で生成できる。
高速真偽判定もできる、こうなると人は馬力で及ばない、
素材には画像や音声、業務課題....と言う AI 化が容易に可能な物と、
道徳、人生目的、福祉政策とかの極抽象語解釈が妥当にできるかの哲学思想道徳言語問題が
あるはずだ。

補足：魂<意識>消滅を追及した仏教。
仏教教えでは人は輪廻転生すると言われる、筆者自身も標準物理⁺論理学からの結論としてあの
世の存在を認めます、こうなると人の真の死が無い事になる。
人は死ねばただの骨と骸骨、
ならば生きてる間に好き勝手し放題、,,,,,秘密結社,骸骨と骨の入社式ソング
海賊船やナチス親衛隊のシンボル☠無神論思想はまちがいになる、コラ-ン教え如く、現世は
来世の為の修行場は正しい、ちなみにナチスは反ユダヤ反キリスト教、隠れて親仏教と言われる。
http://inri.client.jp/hexagon/floorB1F_hss/b1fha200.html#01