УДК 004.9

WORDS PAIRS ANALYSIS BASED ON MULTIDIMENSIONAL LANGUAGE REPRESENTATION

Karpilenko M.S., Veretennikova A.V.
Scientific supervisor – Associate professor Lichargin D.V.

Siberian Federal University

In the work the problem of multidimensional representation of natural language data is considered. The definition of grammatical multidimensional space of natural language words is analyzed. The principle of presenting word pairs is offered. The conclusion about the effectiveness of the presented principle for natural speech generation is made.

Keywords: Natural Language Generation, Semantic Indexes, Language Words and Notions Classification, Turing Test.

The problem of meaningful sentences generation, particularly for the English language, is one of the main tasks of computational linguistics – a discipline important for generalizing the concepts of linguistics, mathematics, computer science, philosophy, psychology, etc for the purpose of natural language formalization. Solving tasks of semantics, discrete mathematics, linguistics and artificial intelligence all in one are aimed to passing the Turing test in more and more difficult conditions considering wide range of words, constructions, facts and expressing the attitude of the speaker.

The solution to these problems is necessary for creating linguistic software, synonymizers, text generators, expert systems, e-learning systems, automatic translation systems, creating natural-language interface.

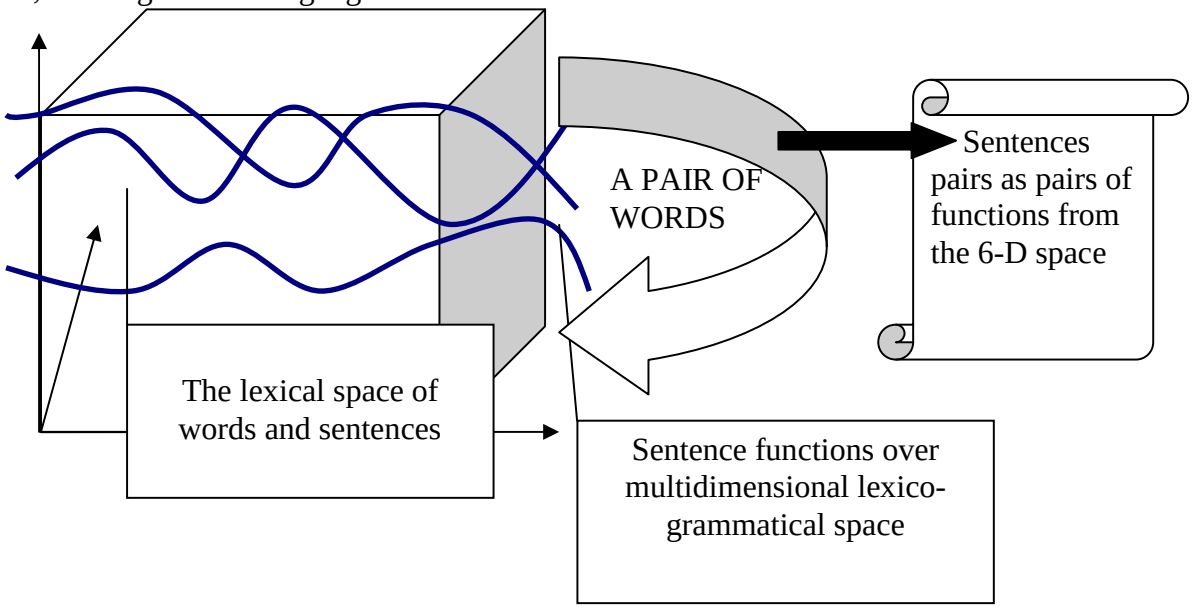

Figure 1. The Lexical and Grammatical Space in the Model of Meaningful Language Generation

The solution of this problem mainly determines solving the problem of passage of the Turing test by software systems, i.e., providing identical dialogue with a person similar to the dialogue with the software system which will be impossible to be distinguished.

The main idea of the paper is to determine the relations of any possible word pairs as a pair of vectors of a multidimensional lexical and grammatical space.

Multidimensional representation of the natural language and determining word pairs relations data is important for constructing linguistic software: electronic interpreters, abstracting systems, expert systems, generative grammars, etc. In this regard, the analysis of a multidimensional model of the language data is relevant at the present stage of the development of information technologies and mathematical science.

The novelty of this work is to show the principle possibility of vector description for word pairs.

First, let's consider multidimensional space of natural language units: words and sentences. Such space of words allows generating grammatically, but not semantically, correct phrases of natural language. Thus, the phrase «See I» doesn't have any grammatical meaning, the phrase «I eat a hat» is grammatically correct, but has no semantic meaning, and phrase «I eat a pear» is both grammatically and semantically meaningful.

Further the examples of using multidimensional coordinates presentation of the language words and notions, which are illustrated on pic.1 in the form of 6-dimensional lexical and grammatical space of the natural language will be given. Each discrete dimension is presented as a classification tree. Pairs of words are considered to be pairs of points in the lexical and grammatical space, and can be presented as pairs of vectors of the 6^{th} dimensional space of English grammar and vocabulary, which is described as follows.

Table 1 – Sections of the Viewed Multidimensional Space as Tablesof Speech Generation by Lulia-Palmer Method

I ***я***	wish to ***желать***	win ***побеждать в***	game ***игра***
we ***мы***	want to ***хотеть***	take part in ***участвовать в***	fair game ***честная игра***
you ***вы***	decided to ***решил***	be a judge in ***судить на***	combat ***поединок***
they ***они***	happen to ***случилось***	lose ***проигрывать***	struggle ***борьба***
	need to ***нуждаться***	take the first place ***занять первое место в***	meeting ***встреча***
	have to ***должен***	conduct ***проводить***	championship ***чемпионат***
	seem to ***кажется***	organize ***устраивать***	contest ***состязание***
		finish ***завершать***	competition ***соревнование***
		open ***открыть***	tournament ***турнир***
		close ***закрыть***	Olympic games ***Олимпийские игры***
		broadcast *брО:дка:ст* ***транслировать***	cup ***кубок***

Thus, in the presentation of the linguistic data it is possible to construct a multidimensional representation of data with the following coordinates of the notions description vector:
gr1 = Parts of speech {«Article», «Adjective», «Noun», «Verb», ...};
gr2 = Member of the sentence {«Determiner», «Attribute», «Subject», «Predicate», ...};
gr3.3.1 = Person {«1st», «2nd», «3rd», «Undetermined»};
gr3.3.2 = Aspect {«Infinitive», «Continuous», «Perfect», «Perfect continuous», «Undetermined»};
g3.1.1 v3.1.2,... – other dimensions, expressed by grammatical categories.
Next, we construct the lexical space of words of the language (data cube) with the following coordinates:
lx1 = Word order {Doer, Action, Recipient, Receiver, Place, Time, Instrument, Method};
lx2 = Topics {Food, Clothes, Body, Building, Group of People, Transport, …};
lx3 = Options for substitution of words in a sentence {to cook, to boil, to roast, to fry, to bake, …, to eat, to chew, …}.

All the grammatical constructions are included into the cells of multidimensional array. The intersection of the vector coordinates, such as, for instance, V[VERB / ATTRIBUTE / PERFECT, ...], defines the cell of multidimensional array with the grammatical construction : «having + VERB-ed». The vector V [ADJECTIVE /, PREDICATE / 1ST PERSON , SUPERLATIVE DEGREE , LONG WORD , ...] defines the construction «am the most + adjective». The relational tables as subsets of that multidimensional array are presented in linguistics in the form of traditional grammatical paradigms.

In table 1 the combinations of words – word pairs are offered. Their examples are given and the pairs of vectors defining the words are presented. The vectors of 3 types are taken into account: gr[A, B, C] is a grammatical space vector, lx[D, E, F] is a lexical space vector and sp[M] is a spelling space vector, which is created by coordinate sp[M[letter1, letter2, letter3, …]] with letters in definite positions in the word as its coordinates.

Table 2 – Possible relations between words from the point of the lexical and grammatical 6-D space

Names of Lexical and Grammatical Relations	**The vector of points in a multidimensional space for word 1**	**The vector of points in a multidimensional space for word 2**	**Examples of such relations**
1. Grammar			
1.1. Difference by a part of speech	gr[«Verb», B, C] + lx[D, E, F]	gr [«Noun», B, C] + lx[D, E, F]	Love – to love
1.2. Difference by a grammatical category	gr [A, B, «Singular»] + lx[D, E, F]	gr [A, B, «Plural»] + lx[D, E, F]	Fan's – fans'
2. Semantics			
2.1. Difference by a topic	gr [A, B, C] + lx[D, «Food, F= «Make»]	gr [A, B, C] + lx[D, «Clothes», F= «Make»]	Cook - sew
2.2. Difference by an object	gr [A, B, C] + lx[D, E, F]	gr [A, B, C] + lx[D, E, F]	Start > launch
2.3. Antonyms	gr [A, B, C] + lx[D, E, F.G.H(disjunction level)]	gr [A, B, C] + lx[D, E, F.G.I (disjunction level)]	To be born – to live – to die – to revive
2.4. Hyperonyms	gr [A, B, C] + lx[D,	gr [A, B, C] + lx[D,	Mother - Parent

	E, F.G.H]	E, F.G]	
2.5. Hyponyms	gr [A, B, C] + lx[D, E, F.G]	gr [A, B, C] + lx[D, E, F.G.H]	Parent - Mother
2.6. Equanims	gr [A, B, C] + lx[D, E, F.G1.H1]	gr [A, B, C] + lx[D, E, F1.G2.H1]	Mother – Father
2.7. Unequanims	gr [A, B, C] + lx[D1 E1, F1]	gr [A, B, C] + lx[D2, E2, F2]	Mother – cup
2.8. Definonims	gr [A, B, C] + lx[D1, E, F]	gr [A, B, C] + lx[D2, E, F]	A cook > to cook > dish > cooked
2.9. Synonym (complete)	gr [A, B, C] + lx[D, E, F.G.H.J.K] + sp[M1]	gr [A, B, C] + lx[D, E, F.G.H.J.K] + sp[M2]	To create – to develop
2.10. Synonym (partial)	gr [A, B, C] + lx[D, E, F.G.H.J.K1]	gr [A, B, C] + lx[D, E, F.G.H.J.K2]	To eat – to have a bite
2.11. Homonyms	gr [A1, B1, C1] + lx[D1, E1, F1] + sp[M1]	gr [A2, B2, C2] + lx[D2, E2, F2] + sp[M1]	Table (furniture) – table (data array)

The group of words {jacket, fur coat, coat, apron, ...} refers to a group of cells of a multidimensional array [NOUN , ? , SINGULAR] + [RECEIVER , THING.CLOTHES, OUTER]. It's hyperonym, a word more general in meaning, will be «clothes» forming the lexical word pair or lexical relation. The chain of such hyperonym relations will be «essence» - «object» - «thing» – «covering» – «clothes». This chain can be presented as a tree of lexical relations. «Something» – «Essence» – («Object» – («Thing» – («Covering» – («Clothes» – («upper clothes», «underwear», «jacket») – «Edible Thing» - («Food» («Fruit» («apple», «crab apple»), - «Drink»))))) – «Action» («Action with Food», «Action with Clothes», …).

There can be pairs and chains of words similar in spelling. For example «bot» - «pot» – «pit» – «pick» according to a number of letters. There will be a tree of graphic similarity, for example, «bot» – («pot» – («pet» – «pit») – («pick» – «pill»)) – («lot» – «let»). The neighborhood of positions in such trees refers to the associative relations of n^{th} degree according to a topical characteristic of the whole tree. Optional association can be presented as a sequence of one-characteristic associations For example, $Association^{optional}$(Professor, Mathematics) := $Association^{hyperonym}$(Professor, Teacher), $Association^{definonym}$(Teacher, Subject, Teach, Student), $Association^{hyperonym}$(Subject, Mathematics). Similaly, $Association^{equanim}$(Mother, Father) := $Association^{hyperonym}$(Parent, Mother), $Association^{hyperonym}$(Parent, Father), Mother≠Father.

The fact that pairs of words can correspond with the pairs of sentences allows us generating the meaningful discourse: Association(Professor, Mathematics) ~ Association(«The Professor is very clever», «But Mathematics is rather difficult»). This is the practical importance of the research.

The method of multidimensional representation of units of the natural language and defining the structures of different levels is a perspective method of the analysis and synthesis of the natural language, as well as the generation of meaningful speech. The proposed classification is effective to generate meaningful speech. Based on the method the relation between words of the natural language can be studied. The relations between the words and word forms are defined as a pair of vectors of 6-D lexical and grammatical space. It can help solving the problem of discourse generation. This principle requires further investigation.