PROCEEDINGS

of the Union of Scientists - Ruse

Book 5

Mathematics, Informatics and Physics

Volume 10, 2013

RUSE

The Ruse Branch of the Union of Scientists in Bulgaria was founded in 1956. Its first Chairman was Prof. Stoyan Petrov. He was followed by Prof. Trifon Georgiev, Prof. Kolyo Vasilev, Prof. Georgi Popov, Prof. Mityo Kanev, Assoc. Prof. Boris Borisov, Prof. Emil Marinov, Prof. Hristo Beloev. The individual members number nearly 300 recognized scientists from Ruse, organized in 13 scientific sections. There are several collective members too – organizations and companies from Ruse, known for their success in the field of science and higher education, or their applied research activities. The activities of the Union of Scientists – Ruse are numerous: scientific, educational and other humanitarian events directly related to hot issues in the development of Ruse region, including its infrastructure, environment, history and future development; commitment to the development of the scientific organizations in Ruse, the professional development and growth of the scientists and the protection of their individual rights.

The Union of Scientists – Ruse (US – Ruse) organizes publishing of scientific and popular informative literature, and since 1998 – the “Proceedings of the Union of Scientists- Ruse".

BOOK 5

"MATHEMATICS, INFORMATICS AND PHYSICS"

VOLUME 10

CONTENTS

BOOK 5

"MATHEMATICS, INFORMATICS AND PHYSICS"

VOLUME 10

This is the jubilee 10-th volume of book 5 Mathematics, Informatics and Physics. The beginning was in Spring, 2001, when the colleagues of the former section Mathematics and Physics decided to start publishing our own book of the Proceedings of the Union of Scientists – Ruse. The first volume included 24 papers. Through the years there have been authors not only from the Angel Kanchev University of Ruse but as well as from universities of Gabrovo, Varna, Veliko Tarnovo and abroad – Russia, Greece and USA.

Since the 6-th volume the preparation and publishing of the papers began to be done in English.

The new 10-th volume of book 5 Mathematics, Informatics and Physics includes papers in Mathematics, Informatics and Information Technologies, Physics and materials from the Scientific Conference 'Information Technologies in Education' (ITE), held at the University of Ruse in November 2012 in the frame of Project 2012-FNSE-02.

web: suruse.uni-ruse.bg

SYSTEM FOR MODELING OF AMBIGUOUS SEMANTICS

Iliyana Raeva

Angel Kanchev University of Ruse

Abstract: *The paper presents a system for linguistics analysis of a natural language text. Based upon a correct syntax analysis, the system generates the semantics of each word, phrase or sentence and solves ambiguities by the selection of appropriately chosen meanings of the ambiguous words selecting the words from semantic nests. The program creates an attributive value matrix for each processed sentence and the relative logical form. The system is to be applied for processing and correcting documents and it can be used as an educational system.*

Keywords*: semantics, logical form, ambiguities, attributive value matrix*

1. INTRODUCTION

The Word Sense Disambiguation (WSD) is a long standing process in computational linguistics. A system solving the problem should be possible to receive as an input undefined text and to select of each word the most appropriate meaning being of acceptable accuracy and efficiency. The most common approach to eliminate ambiguity of the word is the use of the context of the word according to the entered text. For this purpose we use the information for any of the word's significance [4].

2. CONTEMPORARY ALGORITHMS FOR WSD

The algorithms for WSD can be categorized depending on the method, used for the acquisition of knowledge. Here's a sample categorization [2]:

- **<u>Base of knowledge</u>**: WSD using information from a comprehensive dictionary/lexicon or a knowledge base. The dictionary could be computer readable dictionary or encyclopedia (such as LDOCE or WordNet), or manually written. An example is the well known system of McRay;
- **<u>Based on corpora</u>:** WSD using information obtained from the study of several corpuses. This approach could be divided as:

  =<u>tagged on corpora</u>: Information is collected from the corpus the semantic ambiguity of which has been already removed. An example of this is the bilingual corpus of Brown [1] for training statistical WSD algorithms;

  =<u>untagged corpora</u>: Information is collected from the "raw" corpus, from which the semantic ambiguity was not removed.

The modern methods, applying knowledge of the world as selectors (selection preferences) to solve this problem, do not use effectively yet the available knowledge bases. Moreover, their effectiveness decreases when enriching knowledge by increasing the conceptual connections. The effective removement of the ambiguity of the words' meaning requires taking into consideration the dynamic context in processing the sentence in order to find the right set of selectors. In this sense, the system proposed by Kavi Mahesh, Sergej Nirenberg, Stephen Beale (Computing Research Laboratory, New Mexico State University, USA) is such an introductory operator (inference operator), which is the most accurate in the context of WSD Mikrokosmos semantic analyzer.

The method used in the system retains its efficiency even in a high broad knowledge base with a high degree of coupling between the concepts.

3. DESCRIPTION OF THE MODEL – LOGICAL STRUCTURE

The syntax of the language, processing the system, is described by the following contacts free grammar:

L=<T, H, A, R> ,

where
T = {Noun, Name. Pronominal, Adjective, Adverb, Preposition, Verb, Number, Article, λ};
H = {S. VP, PP, NP, NP2, NP3, Art1, Num1, Adj1, Adv1. λ};
A= {S};
R= {
S-> NP, VP
NP -> Art1, Num1, Adj1, Noun
NP-> Name
NP-> Pronominal
VP-> Verb,, Adv1, NP2
NP2->NP3, PP
NP2-> NP3
NP3->NP
NP3->PP
PP-> Preposition NP
Art1-> Art
Art1-> λ
Num1-> Number
Num1->λ
Adj1->Adjective
Adj1-> λ
Adv1->Adverb
Adv1->λ}

The check of the syntax of this context free grammar is implemented by the following preceding matrix:

	Noun	Name	Pron.	Adj.	Prep.	Verb	Adv.	Numb.	Dumb (space)	Art
	0	0	0	0	1	1	0	0	1	0
	0	0	0	0	1	1	0	0	1	0
	0	0	0	0	1	1	0	0	1	0
	1	0	0	0	0	0	0	0	0	0
	1	1	1	1	0	0	0	1	0	1
	1	1	1	1	1	0	1	1	0	1
	1	1	1	1	1	0	0	1	0	1
	1	0	0	1	0	0	0	0	0	0
	0	0	0	0	0	0	0	0	1	0
	1	0	0	1	0	0	0	1	0	0

Although the system processes sentences in which the ambiguous words are nouns, verbs or adjectives, it could be considered as a suitable model for examining the efficiency of the algorithm, since the counted parts of the speech are the majority of the ones in natural language.

The set value function belonging to a semantic nest is defined on the key activities "ambiguous word + test" by a factor of compatibility. This determines the percentage of statistical probability, which is the ratio of the number of word's uses with a specific meaning to the total number of cases in which the words are found with any of their meaning.

As a heuristic criterion, when an **adjective** is considered, the given order follows: first is the noun to be explained, then the verb in the sentence comes and finally, if no longer used as a criterion, the subject of the sentence follows.

For the "criterion" of the **verb** of the sentence, the noun from the following the verb noun phrases is used.

If the ambiguous word is a noun, it is determined by comparing consequentively:

- with its modificator (only with the adjective), if any;
- then with the verb, if it is "to be" this comparison becomes pointless;
- if you have not defined the exact meaning yet and the question concerns the noun from the object noun phrases, the word is juxtaposed with the subject of the sentence.

If after the application of the rules set out above, the ambiguity is not yet removed, the verb and the subject of the previous sentence are used. This is not implemented in the model system by purely technical reasons.

If the result of the comparison is a coefficient 0, it terminates and such an option is considered impossible (meaningless). As a criterion in the demonstration model pronouns are not used, even if they fall within one of the above-described key positions because the anaphoric reference is not processed and therefore undefined data must not be kept.

For example the combination "be" and "blue" is unclear as "he" is undefined in the context phrase.

It is enough to use only the previous sentence of the processed text to remove the remaining uncertainties in the meanings of the words as it in turn uses the previous sentence as well and thus indirectly the whole preceding text is processed to remove the ambiguity. To understand the meaning of a word a larger context of a sentence or two is not necessary. Therefore rarely we have to wait for the next sentence in order fully to understand the meaning of the foregoing one.

Therefore, using words from the following sentences as a criterion to remove ambiguity is unjustified and it unnecessary delays and complicates the algorithm.

The system is self-learning in enriching the vocabulary and the relationships between words, so it is not dependent significantly on the type of the processed text.

The method uses a set of 10 dictionaries in which information for the type of the respective word and for some of its fundamental characteristic parameters is stored (not all of them are used in the system).

They are the following:

1. Adjectives [Category (A 15), Head (A 20)];
2. Adverb [Category (A 15), Head (A 20)];
3. Names [Category (A 15), Head (A 20), Spirit (A 1), Number (A 1)];
4. Nouns [Category (A 15), Head (A 20), Spirit (A 1), Number (A 1), Count(A 1)];
5. Numbers [Category (A 15), Head (A 20), Type (A 1)];
6. Prepositions [Category (A 15), Head (A 20)];
7. Pronominal [Category(A 15), Head(A 20), Type (A 1), Person (N), Number(A 1)];
8. Verbs [Category(A 15), Head(A 20), Tense (A 25), Person (N), Number(A1), Infinitive(A 20)];
9. Meanings [Word (A 25), Meaning(A 50), Category (A 15)];
10. Fuzzy [Word1 (A 30), Meaning1 (A 50), Word 2 (A30), Meaning 2 (A 50), Coefficient (S)],

where:
A is a text type and the number after it indicates how many bytes are allocated to it;
N is an integer;
S is a type (short); t and f (true, false); s and p are numbers (singular, plural); type o and c (ordinary, cardinal); Person (person) 1, 2 and 3;
“Coefficient” shows the percentage of statistical probability.

The total coefficient of variants of the sentence from a given text is calculated as the product of the coefficients of compatibility of the couples word - criterion and an ambiguous word. The percentage of probability of a given option compared with the other versions of the same sentence in the same text is calculated as follows:

in 2 options	in 3 options	in 4 options
x + y = 100 $x=\frac{100a}{a+b}$ $y=\frac{100b}{a+b}$	$\begin{cases} x+y+z=100 \\ \frac{x}{y}=\frac{a}{b} \\ \frac{y}{z}=\frac{b}{c} \end{cases}$ => $\begin{cases} x=\frac{100a}{a+b+c} \\ y=\frac{100b}{a+b+c} \\ z=\frac{100c}{a+b+c} \end{cases}$	$\begin{cases} x+y+z+w=100 \\ \frac{x}{y}=\frac{a}{b} \\ \frac{y}{z}=\frac{b}{c} \\ \frac{z}{w}=\frac{c}{d} \end{cases}$ => $\begin{cases} x=\frac{100a}{a+b+c+d} \\ y=\frac{100b}{a+b+c+d} \\ z=\frac{100c}{a+b+c+d} \\ w=\frac{100d}{a+b+c+d} \end{cases}$

Here x, y, z and w are the values of the probabilities of the respective options in percentages, while a, b, c and d are the common coefficients of the respective options.

4. PROGRAM RESTRICTIONS

Restrictions that are imposed on the type of sentences, processed by the system, are as follows:

* The statements must be in English, be simple, non-negative and grammatically correct.

* There can be only one adjective and / or number, which explains (modifies) the noun and they are before the noun in the stated order.

* There cannot be separate adjectives or numbers.

* A noun cannot be used as a clarification (modifier) of a noun, if an adjective is used it has to be after the noun.

* After the verb maximum two noun phrases could be present.

* Facilitator names and pronouns have no modifiers.

* There cannot be definite articles of own names.

* The verb can be only one word in present or past tense.

The text, processing the system, has the following limitations:

* Each sentence can have maximum two ambiguous words that can be nouns or verbs.

* The system processes no more than five sentences.

* Each ambiguous word can have maximum two meanings.

Input data

The system accepts input text that can be entered manually at startup or by using a standard text file, introduced previously. The program allows the entered text to be saved as a text file with extension * txt. In further processing if unknown words or combinations of words are available, the system requires the introduction of additional data for the corresponding word.

Output data

Output program displays:

• Atributive Value Matrix (AVM) as a tree for each variant of sentence;

• logical form of each variant ;

• the coefficients of consistency of words-criteria's and ambiguous words;

• the common probability of sentence variants providing a possibility for determining the minimum level of probability of the sentence under which the sentence is rejected. If the coefficient of consistency is 1 it is possible to be determined whether in the logical form it will appear as a multiplier and whether the couple of words with this coefficient to be shown.

5. CONCLUSION

The applications could be implemented in systems for correcting and processing of documents (contracts), training systems, expert systems, etc.

To every processing sentence and to every noun phrase a logical form is mapped. Thus, based on the syntax, a certain type of semantic ambiguities are solved, arising from the words in the sentences.

REFERENCES

[1] Peter F. Brown, Vincent J. Della Pietra, Stephen A. Della Pietra, Robert. L. Mercer. "The mathematics of statistical machine translation: parameter estimation" Computational Linguistics, 19(2), 263-311, 1993.

[2] Gatward Richard "A software engineering approach to developing computer", Proceedings SAER'99, 213 – 215, 1999.

[3] Hintikka J. "On semantic Information" In: Hinikka and Suppes, 1970.

[4] Mamdani E. H., J. Efstathion, "An analysis of formal logic as inference mechanisms on expert systems" Intern. J. of Man. Machine Studies 21, 1984.

CONTACT ADDRESS

Pr. Assist. Iliyana Raeva
Department of Mathematics
Faculty of Natural Sciences and Education
Angel Kanchev University of Ruse
8 Studentska Str.,7017 Ruse, Bulgaria
Phone: (++359 82) 888 453
E-mail: iraeva@yahoo.com

СИСТЕМА ЗА МОДЕЛИРАНЕ НА НЕЕДНОЗНАЧНИ СЕМАНТИКИ

Илияна Раева

Русенски университет "Ангел Кънчев"

Резюме: *В статията е представена система за лингвистичен анализ на естестено-езиков текст. На основата на коректен синтактичен анализ, системата задава семантика на отделни думи, фрази или изречения, като решава нееднозначности чрез избор на подходящо избрани значения на нееднозначните думи чрез избор на думите от семантични гнезда.*

Програмата задава атрибутивно-стойностна матрица на всяко обработвано изречение, както и съответната логична форма. Системата е предназначена за обработка и корекция на документи и може да служи като система за обучение

Ключови думи: *семантика, логична форма, нееднозначност, атрибутивно-стойностна матрица*

ISSN 1314-3077

9 771314 307000