Morphology and syntax in a problem of semantic clustering.

D. V. Mikhailov and G. M. Emelyanov

Yaroslav-the-Wise Novgorod State University

An actual *global problem* considered in this paper is the automation of knowledge obtaining about the interaction of semantics, syntax and morphology while establishing the Semantic Equivalence (SE) of Natural Language (NL) texts.

The revelation of SE's class of the sentence is the *major constituent* of computer sense analysis for this sentence. To broadly establish the fact of SE means to prove identity of roles of identical concepts relative to similar situations described by compared texts. Processing of texts on the basis of communicative grammar is closest to the given idea. The search engine Exactus can be a good example here.

Nevertheless, there are tasks of sense's comparison which are different from «query–answer» information retrieval. Example is interpretation of the open form's test task in system of computer-aided testing of knowledge. It is necessary not so much to map the answer to a subject area, as to estimate its affinity to the answer, «correct» from the point of view of a teacher who has created such a test. Here the analysis of the NL-statements's affinity requires taking into account the synonimy described by Standard Lexical Functions, in particular — the Splintered Values and conversives. According to G.S. Osipov, more detailed research of properties of semantic relations in the communicative grammar itself is required. Machine learning's methods here may be employed to study the interaction of semantics, syntax and morphology while establishing SE.

The most actual problem statement for SE with the respect of mentioned requirements to the text comparison consists in the following *(Slide 2)*. Let the set of texts is given. For example, an elements of this set can be a students's developed answers to the test task of open form. It is required: by results of syntactic analysis of initial texts to reveal for each text:

- a set of situations described by this text;
- a set of objects and/or concepts significant in revealed situations;
- a ternary relation, which puts in conformity to each object a situation in which the considered object (or concept) appears concerning the given text.

The *noun's syntactic context* is the basis of revealing *objects* and *situations* here. This context for a noun designating some concept concerning the given situation is a sequence, which consists of predicate word and submitted nouns represented on the *Slide 3*. The role of object concerning the situation designated by a predicate word is defined by the type of submission relation between this word and the word to the right of it in a sequence. It can be determined by the case of dependent word and by preposition, which connects the syntactically main and dependent word. Submission relation's transitivity which follows from sense correlation of submitted words allows to state that any noun of sequence designates some object which is significant in given situation.

On the basis of revealed correspondences between objects, situations and roles an initial texts are grouping on the basis of objects occurrence in similar situations. It is most natural to decide the given problem of semantic clustering by involving the methods of Formal Concept Analysis (FCA, *Slide 4*). At this case to

the ternary relation between the sets of texts, objects and situations the formal context is put in conformity and the Formal Concept Lattice for initial set of texts is under construction. Thus the sense affinity's analysis of texts is reduced to investigation of the lattice's qualitative characteristics. Lattice's visualization by means of a line diagram allows to represent graphically the text grouping.

Nevertheless, the *problem of accuracy of syntactic analysis* as a toolbox to reveal objects and attributes is *actual* here especially for Russian. Known syntactic analyzers, in particular, Cognitive Dwarf, release parse strategy on the basis of the most probable relations. At the same time, often it is required to investigate the nature of revealed syntactic relations. Respecting the features of language situation's representation at the case of incorrect parse it is necessary to analyze the cause of usage of one or another syntactic strategy (or rule).

The purpose of this paper is to develop a *mathematical model* for revelation and classification of the most probable syntactic relations on the set of Semantic Equivalent phrases.

The offered decision of a problem is based on laws of sense expression in given Natural Language by its informant. The idea of dividing a language experience of human is basic here. This experience can be divided according to division of a conceptual picture of the world. We can consider usage situation of Natural Language as a basis of its genesis.

A *situation of Natural Language usage* in our understanding is the description of new social experience (the content of joint actions) by means of this NL. This description is carried out in some character system and have a purpose to generalize and share human knowledge. On the *Slide 5* a formal model of language context accumulated by some such situation is represented. This model can be represented by a triple including a set of objects which participated in situation, a set of relations between objects and a set of description forms for situation in the given Natural Language.

Let's assume, that the situation of Natural Language usage is given by a set of synonymic phrases and each of which describe the same reality situation concerning the given language context. Here a choice of phrase for situation's description is equiprobable. By virtue of arbitrariness of relations between an objects in situation let's assume, that the set of mentioned relations consists of syntactic relations between the words, which designate objects at the set of synonymic phrases. Thus the set of objects can be considered as a set of concepts significant in the given situation of reality. This set include the verbal designations of this concepts (including the designation for given situation).

So we have the next *problem. It is given* a set of synonymic NL-phrases. *It is required to reveal* a relations between the designations of concepts significant in described reality situation by using the given relations as an attributes for words concerning the given situation of Natural Language usage.

Let's consider a text from the point of view of symbols which make it. Then for any text from the given synonymic set it is fair to select *(Slide 6)* an inflectional part and some invariant part common for all texts. By means of inflectional part the syntagmatic dependences are expressed. Syntagmatic dependences define linear coexistence of wordforms and are set by syntactic relations.

For syntactic relation's establishment an *inflexion* is significant. This is a word's

part which is altered through declension or conjugation and located at the end of wordform. On the basis of inflections combinations a morphological dependences can be revealed. Because the shown dependences are one of *realization's* ways for *syntactic relation* then the syntactic relation itself can be revealed by pairwise comparison of alphabetic structure of different words with revelation of invariant part and inflectional part.

Let's enter into consideration an *index set* for invariant parts of all words used in all phrases from the given synonymic set. Let's call a sequence of indexes of invariant parts of words presented in the given NL-phrase as a *linear structure's model* for this phrase *(Slide 7)*. Here an order of indexes in a model corresponds to the sequence of respective words in a phrase. Therefore a linear structure's model allows to restore unequivocally the NL-phrase on the set of all words for all phrases from the given synonymic set. And vice versa, for any phrase from the given synonymic set it is possible to construct its model unequivocally on the set of indexes.

For forming a set of syntactic relations concerning the given situation of Natural Language usage *it is necessary to find* the set of mentioned models, each of which corresponds to *requirements of projectivity*. A *linear structure's model* for NL-phrase should be considered as *projective* in the substitutional plan if all arrows for revealed syntactical links can be drawn without crossings by one side from the line along which the model is written. Besides, if from the position of some index it is directed more then one arrow then this position can not be covered by arrows directed from the positions of other indexes.

Taking into account the *linear nature of syntagmas* let's supplement the mentioned projectivity's restrictions as follows. Let's assume, that *linear structure's model* for NL-phrase *is projective* relative to the set of syntactic relations at the given situation of Natural Language usage if the summary length of all links concerning the model dos not exceed its length. Here a pair of indexes relative to which a link is set corresponds to the single *syntagma*. *The link* is *acceptable for the linear structure's model* of some phrase if the given synonymic set contains a pair of phrases for which their models include as a subsequence the pair of indexes related to the considering link, or include this index pair but written back to front.

By means of grouping an indexes pairs concerning which the links for linear structures's models are set a *graph* of *syntagmas* is formed, see *Slide 8*. On the basis of this graph the *syntactic precedent tree* for the given synonymic set is build. Using the paths in this tree a laws of linear coexistence of inflections can be revealed on the basis of the *formal context of inflectional compatibility* represented on the *Slide 8*. Here a revealed classes of syntactic relations are correspond to the alteration type for inflectional part of dependent word. An investigated problem of syntactic analysis's accuracy is a typical for situations with two or more participants. Therefore let's assume that the quantity of child nodes for root is greater than one.

Now let's consider a problem of inflexions's revelation for words in structure of Splintered Values and conversives. We shall consider a Splintered Predicative Value (SPV) as a community of auxiliary verb (a copula) and some noun denoting a situation *(Slide 9)*. The invariant part of SPVs cannot be found in all NL-phrases of the given synonymic set. It is fair to approve the same about conversives — a

words designating the same situation from the point of view of different participants. On the *Slide 10* the properties of *linear structure's model* of NL-phrase which are actual for searching a place of *unrecognized predicate word* in a structure of *syntactic precedent tree* are represented. The *basic* for *proof* of this properties is the assumption about the quantity of participants of situation designated by a predicate word with the respect of projectivity of linear structure's model. For words which are a members of SPV according to *Lemma 2* an invariant and inflectional parts can be revealed by comparison of their alphabetic structure with all words (not necessary a members of SPVs) invariant part of which is unrecognized and which contain in NL-phrases not satisfying a condition of *Theorem 1*. The prevalence of likenesses in alphabetic structure of compared words is the necessary condition to allocate an inflectional part *(Slide 11)*. With the respect of branch direction typical for SPVs in submission tree a *syntactic precedent tree* which was initially revealed will be transformed taking into account new indexes for words within a structure of SPVs according to rules represented on the *Slide 11*. As a result the *formal context of inflectional compatibility* will be formed on the basis of those NL-phrases which most fully describe the given situation of reality concerning a considering language context.

The offered model for revelation and classification of syntactic relations has been approved at computational experiment on a material of results for open form test *(Slide 12)*. Here a basis of forming the *formal context of inflectional compatibility* is made by NL-phrases with maximal projectivity and minimal quantity of words without prototypes in alphabetic structure of invariant part *(Slide 13)*. On the *Slide 14* a Formal Concept Lattice for the *formal context of inflectional compatibility* for resultant set of NL-phrases is represented. Here a sense interpretation of the lattice can be obtained by revelation of morphological classes of words on the basis of Duquenne-Guigues set of implications *(Slide 15)* for the considering formal context with the respect of structure of *noun's syntactic context* according to rules represented on the *Slide 16*. A syntactic relations themselves can be revealed by analysis of supremum for each pair of Formal Concepts in a lattice. On the basis of affinity of inflection type for dependent word we can reveal classes for those relations. An area in a lattice corresponds to the separate class and the Least Common Superconcept of this area corresponds to the class precedent. In an example on the *Slide 14* relations classes correspond to the inflection of Russian adjectives (нежелательн-ого, эмпирическ-ого) and nouns within a structure of Russian Genitive Constructions (результат-ом переобучени-я, следстви-ем переобучени-я). By virtue of transitivity of syntactic relation within the frameworks of sequences of submitted words a class for nouns within a Genitive Constructions can include combinations of nouns with verbs outside these constructions.

The basis of lattice forming are those NL-phrases which describe the situation most exactly and therefore more accurately express the sense. Therefore morphological dependences revealed on the basis of affinity of inflection type for dependent words correspond to the most probable syntactic relations for the language description of the given situation.

The proposed model for revelation of laws of wordforms's linear coexistence allows to decide *two* important *tasks* which *are actual for semantic clustering* of

NL-texts.

Firstly, *to reveal* automatically *the best way to express some thought in the given Natural Language*. That allows to minimize the errors of syntactic analysis at its usage for revealing objects and attributes.

Secondly, *to automatize the development of strategies and rules of syntactic analysis*. It is *especially actual* at investigation of implementation cases for given grammatical patterns in subject-oriented text corpora. An appraisal of forming knowledge here are based on similarity measures for lattices by analogy with similarity measures between the Formal Concepts.

In this paper we have considered classes of relations for words with a changeable part in the end of wordform. Nevertheless, it is extremely interesting *to develop the offered clustering method* with reference to alterations in structure of a word's stem. Here it is necessary to note unstable vowels, vowel gradations and interchanges of consonants in a stem structure, and also alternative forms of stems. In particular, a special consideration must be given to inclusion of numerals into syntactic context of Russian noun. In Russian the phenomenon of alternation for numerals is especially actual. For example: «триста», «трехсот», «тремстам», «триста», «тремястами», «трехстах». In this connection another *important direction of the further researches* is recognition of paronymic words within a structure of synonymic phrases. The given research will give the most fruitful results together with quantitative studying of variability at level of morphemes and Russian lexemes.