Computer text understanding in a general problem of pattern recognition.

D. V. Mikhailov, G. M. Emelyanov

Yaroslav-the-Wise Novgorod State University

The Semantic Equivalence (SE)'s class's revelation is a major constituent of computer text understanding. From the point of view of pattern recognition an understanding a Natural Language (NL)'s text means to correlate this text with one of known semantic standards with estimation of affinity's degree.

The offered work is devoted to the three most important tasks which are represented on the *Poster 1* and be necessary to decide a problem of texts's semantic affinity's estimation:

- semantic standard's forming. Finally this standard corresponds to a class of semantic equivalence;
- the structure of the semantic standards's database (DB) and filling this DB for a given subject area;
- the quantitative measure of NL's statements's semantic affinity concerning the standards's DB.

Let's consider a usage situation for natural language as a basis of semantic standard's forming. In substantial plan a usage situation for natural language is the description of some reality fact by means of the set of semantically equivalent NL-phrases. Owing to this conception each usage situation for natural language can be identified with the situation of semantic equivalence in the given natural language. An arbitrariness of description forms for natural language's usage's situation gives a possibility to use the trees of syntactic submission as a forms of situation's description. So a precedent of SE's class (i. e. a semantic standard) can be formed by results of syntactic analysis of each SE-phrase from the given set. Because to establish the fact of SE means to prove identity of roles of identical concepts relative to similar situations described by compared texts the set of stems of the words syntactically submitted to other words will be the most comprehensible variant of set of objects for a standard class. Thus attributes will characterize a type of relation between the main and dependent word which is defined by combinations of stems and inflections of dependent and main word, and also «stem–inflection» combinations for a dependent and main word, accordingly.

The specified relations between objects and attributes of semantic equivalence's class define a set of characteristic functions which specify a sense of each NL-phrase of semantically equivalent concerning the given natural language's usage's situation. Let's note that sense's characteristic functions must be considered as the relationship which directly defining the data in a semantic standards's database. For formation and clustering of these relations a language context of NL-usage's situation can be described by means of the formal context used in the theory of formal concept analysis (FCA) and mapped on the *Poster 2*. Thus to objects's set of standard class a set G is put in conformity. And, analogously, a set M in the shown formal context is corresponded to attributes's set of standard class. Note that the FCA as an expansion of the lattice theory is the tool of conceptual clustering because formal concepts in a lattice are classes with interpretation in the form of the concepts's intents. Thus, as well as in a classical problem of pattern recognition, a revealed classes of semantic equivalence differ in abstraction's degree which depends on usage frequency for main words of analyzed combinations in various syntactic contexts.

Let's note that for estimation of affinity of NL-statement to the standard a classes of one abstraction level are significant. These classes are correspond to submission of nouns designating participants of the situation to those words which name the situation and don't enter into Splintered Predicative Values (SPV). In our understanding, each SPV consists of auxiliary verb (a copula) and some noun denoting a situation. The rule of exception of objects and attributes of the splintered predicative values is represented on the *Poster 3*. This rule evidently follows from the properties of Duquenne-Guigues set of implications for formal context representing a usage situation for natural language.

After removal of splintered predicative values's information the formal context for a usage situation for natural language reflects classes of semantic relations defined by roles of participants of the described situation of reality concerning this situation itself. Nevertheless, the number of language description forms for the NL-usage situation initially doesn't specified. Actually it means that synonymic words can designate concepts with various degree of abstraction. In practice this degree the more if more the number of situations concerning which the concept appears in some fixed role. The specified fact can be considered as a basis of definition of affinity's measure for the NL-usage situations generated independently from each other. At this case the affinity measure itself is based on the lattice representation of situation as information unit of subject area's thesaurus. That means the thesaurus can be represented by formal context as shown on the *Poster 4*. The objects of this formal context are correspond to the natural language's usage's situations for given subject area. Into attributes's set of thesaurus's formal context should be included attributes of a formal context for each situation in aggregate with symbolic pointers to objects of formal contexts of separate situations, «stem–inflection» communications for syntactically dependent words and combinations of stems of dependent and main words. Let's assume that the information of the splintered predicative values is removed from sets of objects and attributes of each situation in advance.

On the *Poster 5* the formal definition of the affinity relation between the NL-usage situations is given. This definition is based on the introduced lattice model of thesaurus. The given relation will take place if for each object within the frameworks of NL-usage situation's formal context for analyzed statement in a formal context of standard there will be an object-prototype (it corresponds to synonymic word) which is characterized by similarity of inflectional and lexical compatibility. There the specified kinds of compatibility are considered either concerning only of formal contexts of standard and analyzed situation of NL-usage (see *Condition 1*), or with attraction of the thesaurus's formal context (see *Conditions 2–4))*). Considered definition of affinity of NL-usage situations reflects cases of synonyms among words, main relatively to compared words (see *Condition 2) and 3))*, including taking into account a generic relations (see *Condition 4))* and, hence, respects the degree of abstraction of the concepts designated by synonymic words. Here the analysis of situations's affinity includes comparison of sequences of two and more co-ordinated words. The Russian's example: *«средняя ошибка на обучающей выборке»*⇔*«эмпирический риск»*. Synonymic transformations of NL-phrases does not change a structure of such sequences. Satisfiability of the *Definition 4* is analyzed only for main words (in considered example they are *«ошибка»* and *«риск»*). Sequences are considered as mutually replaceable if it is possible to construct them on a formal context of the thesaurus by a set of attributes of a kind «главное-основа:» for the same NL-usage situation. Thus the main words of sequences should submit equally to the same word what can be checked on a combination of inflections.

For NL-usage situations formal contexts of which are satisfy the *Definition 4*, the affinity measure is calculated according to the *formula (5)* from presented on the *Poster 6*. In our understanding, the measure of situations's affinity will be defined by number of attributes common for objects of compared situations concerning a formal context of the thesaurus. Actually the more words can be syntactically main relatively to each word of the compared pair, the more value for the affinity measure will be. If at least for one object in structure of a formal context of analyzed situation there are no feasible conditions of *Definition 4*, the measure of affinity to a standard is assumed equal to zero.

As an example, let us consider the description of relation's presence between *overfitting* and *empirical risk*. Some facts of subject area «Mathematical methods of learning by precedents», which were used for thesaurus formation, are presented in the *Table 1* on the *Poster 7*. Let the obviously correct (i. e. «standard») description of relation between overfitting and empirical risk is described by four synonymous simple extended Russian sentences, see *Table 2* on the *Poster 8*. Sentences 1 and 2: *«Переобучение (=переподгонка) приводит к заниженности эмпирического риска»*. Sentences 3 and 4: *«Заниженность эмпирического риска связана с переподгонкой (=переобучением)»*. Let's assume that we have four analyzed variants of NL-usage situation. The first three of them are in the affinity relation with the standard according to the *Definition 4* and describe the same relation between overfitting and empirical risk, but by means of one sentence. The first variant: *«Заниженность средней ошибки на обучающей выборке связана с переобучением»*. The second variant: *«Заниженность средней ошибки на обучающей выборке связана с переподгонкой»*. The third variant: *«Переобучение приводит к заниженности средней ошибки на обучающей выборке»*. The fourth variant not only doesn't describe the considered fact, but also is incorrect from the point of view of the correspondent subject area: *«Заниженность средней ошибки на обучающей выборке приводит к эмпирическому риску»*. Performing the syntactical analysis by «Cognitive Dwarf» software (http://cs.isa.ru:10000/dwarf), we identify the stems, inflections, their combinations and obtain the formal contexts for standard and analyzed NL-usage situations. Apparently from the *Table 3* on the *Poster 9*, the greatest value of affinity to the standard has the *Variant 1* of analyzed situations. This is caused by that the attributes of objects of the formal context for this variant are shared by a greater quantity of objects of the standard's formal context, than attributes which are available for objects in *Variants 2* and *3*. In other words, attributes for the *Variant 1* are more stereotypical relative to standard than attributes for other variants. For the *Variant 4*, as one would expect, the value of affinity to standard is equal to zero.

In conclusion let's note, that the offered model of thesaurus allows to reduce the size of standards's database and retrieval time in it owing to hierarchy of information's representation. Thus information compression will be more if more relevant to the given subject area will be each fact description presented in a lattice. Quantitative estimations of NL-description's coverage of subject knowledge in a thesaurus lattice is a topic for separate applied research.