LOSSLESS-IN-SENSE TEXTUAL INFORMATION'S COMPRESSION BASED ON KNOWLEDGE BASE ABOUT SYNONYMY[1]

D. Mikhailov, G. Emelyanov [2]

[2] Yaroslav-the-Wise Novgorod State University,
173003, Russia, Velikii Novgorod, ul. Bol'shaya St. Petersburgskaya, 41, tel.: (8162)627940
e-mail: Dmitry.Mikhaylov@novsu.ru, Gennady.Emelyanov@novsu.ru

This paper is devoted to the problem of textual information's transmission with minimal loss of useful component. This problem's decision offered is to consider a situation of using of natural language as a unit of formalized description of text semantics with application of methods of Formal Concept Analysis.

Introduction

In the present tense one of the main goals of artificial intelligence is to accumulate subject-oriented knowledge and to provide their exchange among people. Not small importance role here is played by knowledge represented in Natural Language (NL) texts.

For example, to interpret a result of executing an open-form test the system of computer-aided testing of knowledge must take into account different NL-description forms given by different experts for the same reality fact using the same natural language. The problem to seek a most rational plan for sense's transfer is actual here. The sense as a result has to be reflected in a maximum compact volume of text data. These data participate in estimation of affinity to the given correct variant for the trainee's answer. The current paper represents the solution of mentioned problem on the basis of NL-usage's situation's sense-standard's conception offered by authors.

Situation of NL-usage

Let Ts be a set of Semantically Equivalent (SE) NL-phrases which are various forms of description of some reality fact. These SE-phrases define the NL-usage's situation. Let's represent a single situation of NL-usage by a triple:

$$K = (G, M, .I), \qquad (1)$$

named a Formal Context (FC) in the theory of Formal Concept Analysis [1]. Here an objects's set G consists of stems of those words which are syntactically submitted to any other words from SE-phrases that are an elements of Ts. An attributes's set M include attributes which point to stems and inflections of words syntactically main to words with stems from G. In M a «stem–inflection» relations for syntactically main word and combinations of inflections of dependent and main words are represented also.

The problem statement: to form the relation $I \subseteq G \times M$ by analysis of symbolic structure of phrases from Ts. Forming the set I must be based on those phrases which meet the requirement of sense's compact expression.

Sense standard and its application

In tasks of classification the compactness hypothesis is understood as the presupposition that similar objects lies in same class more frequently than in different ones [2]. If the sense of phrases from the set $\{Ts_i : Ts_i \in Ts\}$ can be represented as a set of functions which relate concepts designated by words then each such function:

– is given on the set of symbolic chains which correspond to the stems of words from phrases $Ts_i \in Ts$;

– has values's range which is unambiguously determined by some $I' \subset I$.

A compact representation of sense here means to minimize the symbolic length of Ts_i at maximization of number of words $w_j \in Ts_i$

[1] This work is financially supported by RFBR (project №13-01-00055).

which are most generally used in different phrases from Ts (taking possible synonyms into account).

Let's designate further an index set for invariant parts (they associate with stems) of words of phrases from Ts as J. The ordered sequence of such indexes for some $Ts_i \in Ts$ we'll name as Model of its Linear Structure (MLS), $Ls(Ts_i)$.

Let LS be a set of linear structure's models given on J for phrases from Ts.

Lemma 1. A pair $\{j_1, j_2\} \subset J$ corresponds to synonyms if $\exists \{Ls(Ts_1), Ls(Ts_2)\} \subseteq LS$:

$$\begin{cases} Ls(Ts_1) = J_{bef} \bullet \{j_1\} \bullet J_{aft} \\ Ls(Ts_2) = J_{bef} \bullet \{j_2\} \bullet J_{aft} \end{cases},$$

where $J_{bef} \subset J$, $J_{aft} \subset J$ and "$\bullet$" is the concatenation operation defined on J.

Let PJ be the set of pairs meeting the condition of Lemma 1. Let's replace indexes which are members of pairs from PJ, by some $j \in (\mathrm{N} \setminus J)$ in all models from LS. The transformed set LS let's designate as LS'.

Statement 1. Let $\{J_1, J_2\}$ be a pair of sequences of indexes in $Ls(Ts_i)$, where $J_1 = \{j_1^1, \dots, j_2^1\}$, $J_2 = \{j_1^2, \dots, j_2^2\}$, and both (j_1^1, j_2^1) and (j_1^2, j_2^2) correspond to the syntactic links. Thus the sense standard for NL-usage situation is defined by those $Ts_i \in Ts$, in MLS of which

$$\left[\begin{array}{l} J_1 \subset J_2 \\ J_2 \subset J_1 \\ |\,J_1 \cap J_2| = 1 \\ J_1 \cap J_2 = \varnothing \end{array}\right., \qquad (2)$$

and summary length of mentioned sequences for all links revealed on Ts_i is minimal.

Statement 2. Let $fr(w_j)$ be a frequency of occurrence of word w_j (independently of its form) in all $Ts_i \in Ts$. Thus basis of standard are made the phrases with maximum of words entering into special cluster $Clust$:

- the word with a maximal value of this frequency will be a member of $Clust$;
- for $\forall \{w_j, w_k\} \subset Clust$ and $\forall\, w_l \notin Clust$ is true that

$$\begin{cases} |fr(w_j) - fr(w_k)| < |fr(w_j) - fr(w_l)| \\ |fr(w_j) - fr(w_k)| < |fr(w_k) - fr(w_l)| \end{cases}.$$

Remark. At formation of $Clust$ a possible synonyms for analyzed words (according to Lemma 1) are taking into account, therefore for any w_j it is more correct to estimate the value of $fr(w_j)$ concerning LS'.

Let $J_{Cl} \subset J$ be a set of indexes of words entering into $Clust$. Let's consider a set

$$\begin{aligned} & LC = \textstyle\bigcup_i LS_i : LS_i \subset LS, \exists\, Ts_i, Ts_j \in Ts: \\ & \qquad Ls(Ts_i) \in LS_i, \\ & \qquad |Ls(Ts_i) \cap J_{Cl}| \to \max, \\ & \qquad ((Ls(Ts_j) \in LS_i) \wedge (Ts_j \neq Ts_i)) \to \\ & \qquad\qquad \to (Ls(Ts_i) \cap J_{Cl}) \subset Ls(Ts_j). \end{aligned}$$

As follows from Statement 2, sense standard is defined by those phrases, linear structure's models of which are members of LC.

For forming the attributes's set for NL-usage's situation's sense's standard in a form of FC (2) it is necessary to find index pairs which satisfy the condition (2) and to define the direction of syntactic link for each pair.

Algorithm 1. Forming the links.

```
Input: $LS$;
Output: $R_J = \{((j,k), Dir) : Dir \in \{\leftarrow, \rightarrow\}\}$;
Begin
1: $R_J := \varnothing$;
2: forming $LC$ on the basis of $LS$;
3: for all $Ls(Ts_i) \in LC$
4:   $P_i := \{(j,k) : j,k \in Ls(Ts_i), j \neq k\}$;
5: $P := \bigcup_i P_i$ taking $(j,k) \Leftrightarrow (k.j)$ into account;
6: $P' := \{(j,k) \in P : frq((j,k), LS) > 1\}$;
7: for all $(j,k) \in P'$
8:   if $Dir(j,k)$ is found then
9:     $R_J := R_J \cup \{(j,k), Dir\}$;
End {Algorithm 1}.
```

Here $frq((j,k), LS)$ is the frequency of occurrence of the pair (j,k) in the models from LS taking into account that $(j,k) \Leftrightarrow (k.j)$.

For each pair (j,k) from revealed on Step 6 of Algorithm 1 there are three stages to find $Dir(j,k)$. The first stage is the checking for falsity of the link corresponding to pair.

[1] This work is financially supported by RFBR (project №13-01-00055).

Definition 1. Let $\{j,k,l\} \subset J$, and to indexes j, k and l the words's stems $St(j)$, $St(k)$ and $St(l)$ correspond. The link associated with (j,k) is identified as false relatively to the given NL-usage's situation at simultaneous fulfillment of next conditions:

1. $\exists\, Ts_i \in Ts:\ j,k,l \in Ls(Ts_i)$.
2. In the given subject area can be found NL-usage's situation where link between $St(j)$ and $St(k)$ was identified as false, but the link either between $St(j)$ and $St(l)$, or between $St(k)$ and $St(l)$ exists.

Remark. Initial system knowledge about true and false links are formed in a mode of interview with expert. The cumulative knowledge related to specific NL-usage's situation corresponds to the boolean vector

$$\left(d_1,\ldots,d_k,\bar{d}_{k+1},\ldots,\bar{d}_n\right),$$

where $d_1,\ldots,d_k$ are identified with the true links, and $\bar{d}_{k+1},\ldots,\bar{d}_n$ – with the false links.

A pair (j,k) will be checked on a possibility of identification with the links revealed earlier if there is no proof for its identity with any known false link.

Let $w(j) \in Ts_i:\ w(j) = St(j) \bullet Fl(j)$, where the symbolic chain $Fl(j)$ represents an inflection of word $w(j)$, and by symbol «•» the operation of concatenation is designated. Similarly, let $w(k) \in Ts_i$ and at this case $w(k) = St(k) \bullet Fl(k)$. Let's designate a set of links revealed earlier, as Lnk. Each element of Lnk is represented by quadruple

$$(Id, St_1, St_2, FCm),$$

where Id is an ID number of NL-usage's situation; St_1 and St_2 are a stems of main and dependent words, respectively; FCm is a list of pairs of the form «main word's inflection – dependent word's inflection».

A pair (j,k) is put in conformity of link $((j,k),\rightarrow)$ concerning the given NL-usage's situation if for some other with ID number Id

$$\exists\,(Id, St_1, St_2, FCm) \in Lnk:\ St(j) = St_1,$$
$$St(k) = St_2, \text{ and } (Fl(j), Fl(k)) \in FCm.$$

In a case when $St(j) = St_2$, $St(k) = St_1$, and FCm contains a pair $(Fl(k), Fl(j))$, the pair (j,k) will correspond the link $((j,k),\leftarrow)$.

As well as at a stage of formation of initial knowledge, a pair (j,k) will be checked with attraction of expert's interviewing, if there are no identification for this pair with any link revealed earlier (neither true nor false links).

Using the found R_J, further there is a selection of phrases $Ts_i \in Ts$ to form the attributes's set for NL-usage's situation's standard represented by the FC (1).

The first step from $\forall\, LS_i \subset LC$ eliminates such linear structures's models, that include indexes which haven't been entered into any link in R_J. Let's designate the set LC transformed by this way, as LC^*, similarly, $\forall\, LS_i \subset LC$ – as LS_i^*.

For each $LS_i^* \subset LC^*$ one needs to select Ts_i:

$$Ls(Ts_i) \in LS_i^*,\ |Ts_i| \to \min. \qquad (3)$$

Let's designate as Ts^* a set of phrases $Ts_i \in Ts$ meeting the condition (3).

Final step of forming the FC of a kind (1) for NL-usage's situation's standard consists in formation of attributes's set M and relating it to objects's set within the frameworks of $I \subseteq G \times M$ on the basis of found R_J and Ts^*.

With the purpose of more exact revelation of standard's objects and attributes the procedure which co-ordinates knowledge concerning different situations of NL-usage on a given subject area, is introduced. Let model (1) be a unit of thesaurus represented by a triple

$$Kth = (Gth, Mth, Ith), \qquad (4)$$

where Gth consists of labels of individual NL-usage's situations; Mth includes the attributes of FCs (1) for all $gth \in Gth$. Also for each $gth \in Gth$ Mth contains indications to objects of its FC (1), combinations «stem–inflection» for dependent, and combinations of stems of dependent and main words. Model (4) allows to define the procedure of knowledge units's coordination, using the following rule.

Rule 1. Let St_j be the stem and Fl_j be the inflection of word w relatively to NL-usage's situation S_j. Let's suppose that $w = St_1 \bullet Fl_1$ for NL-usage's situation S_1 and $w = St_2 \bullet Fl_2$ – for S_2, at that $St_1 = St_2 \bullet suf$, where suf contains one symbol as minimum. Then

[1] This work is financially supported by RFBR (project №13-01-00055).

concerning S_1 the stem St_1 can be replaced to St_2, and inflection Fl_1 – to $Fl_3 = suf \bullet Fl_2$ only if the occurrence of Fl_3 and Fl_2 in relations from $Ith \subseteq Gth \times Mth$ won't decrease at fulfillment of these changes.

So, for NL-usage's situations shown in Table 1, coordination of their standards as the units of thesaurus, described by a model (2), gives the additional decrease of its size in average by 1,5%. For comparison, Table 2 shows numbers of SE-phrases defining NL-usage's situation (N_1) and standard (N_2), initial numbers of objects (N_3) and attributes of NL-usage's situation (N_4), numbers of objects (N_5) and attributes of standard (N_6).

Table 2. NL-usage's situations's standards

i	1	2	3	4	5	6
N_1	56	28	29	30	6	10
N_2	8	9	7	9	1	2
N_3	18	17	15	13	12	14
N_4	177	186	173	162	94	81
N_5	9	12	12	11	8	12
N_6	82	90	80	69	35	53

Table 1. Russian language's usage's situations having standards, presented in Table 2

i	Phrase of the maximal length from defining NL-usage's situation
1	*Нежелательное переобучение является причиной заниженности средней величины ошибки алгоритма на обучающей выборке.*
2	*Тренировочная выборка, на ней проявляется эффект заниженных значений средней ошибки, причиной же является переусложненная модель.*
3	*Контрольная выборка, принятие деревом решения на ней будет с большей вероятностью ошибки именно по причине переподгонки.*
4	*Оценка частоты ошибок на выборке, взятой в качестве контрольной, может для алгоритма оказаться заниженной по причине переподгонки.*
5	*Заниженность оценки ошибки распознавания зависит от выбора правила принятия решений.*
6	*Число закономерностей алгоритмической композиции влияет на частоту ошибок логического классификационного алгоритма на контрольной выборке.*

Table 3. Estimating the amount of memory for storing NL-phrase

i	1	2	3	4	5	6
n	12	15	16	17	10	14
$vol(n)$	$4.790 \cdot 10^8$	$1.308 \cdot 10^{12}$	$2.092 \cdot 10^{13}$	$3.557 \cdot 10^{14}$	$3.629 \cdot 10^6$	$8.718 \cdot 10^{10}$
$vol_1(n)$	648	795	416	442	20	42
$vol_2(n)$	168	225	80	187	20	42

Thanks to offered idea of NL-usage's situation is possible to estimate the amount of memory for texts's storing. Usually for phrase consisting of n words the value $vol(n) = n!$ is taken here. Using the standard of NL-usage's situation here allows to give the upper estimation as $vol_1(n) = l_1 \cdot n$ and lower – as $vol_2(n) = l_2 \cdot n$, where l_1 and l_2 are a numbers of SE-phrases defining NL-usage's situation and its standard. Comparison of such estimations is presented in Table 3 for NL-usage's situations from Table 1.

Conclusion

The offered method of revelation NL-usage's situation's standard is implemented in demo-release of knowledge-control system presented in [2] with source code on Visual Prolog 5.2 (see section «Participant: Dmitry.Mikhaylov» of «Pages of participants»). Let's note that knowledge coordination according to Rule 1 is similar to self-organization of words sense in multi-agent approach [3], what allows to find words's sensible fragments's co-occurrence's dependences's systems using a model (1), and reduce a search at context's model's forming.

References

1. Ganter B. and Wille R. Formal Concept Analysis - Mathematical Foundations // Berlin: Springer-Verlag, 1999.
2. http://www.machinelearning.ru (in Russian, access date: 27.04.2013).
3. Minakov I.A. Integration of the professional knowledge presented in the form of natural language texts // Herald of SSTU, series «Technical sciences». Samara, 2007. № 1(19). P. 28–35 (in Russian).

[1] This work is financially supported by RFBR (project №13-01-00055).