CRITERIA OF LOAN WORDS IDENTIFICATION

Alla Zaboleeva-Zotova, Ilya Prokhorov

Abstract: This paper describes the criteria of words relation degree identification based on main adoption methods.

Keywords: Criteria, adoption methods, loaned words.

ACM Classification Keywords: I.6 Simulation and modeling - Model Development

The development of the modern languages is continuing in the present time. One of the important ways of this development is the loaning of the words. It was being used for different reasons. The basic one is the absence of a suitable word for some concept or object name in the language. But the essential one is the influence of the fashion. For example, the words *корсаж, пальто, буфет, салон, мебель, туалет, бульон* and *котлета* appeared in the Russian language due to the fashion on French language during the reign of the tsar Peter I and later in the end of the 18th – beginning of the 19th centuries.

The identification of the source and the method of word loaning are defined during etymological analysis. This problem has a high dimension even in the case of the direct source identification. The situation is even more complex considering that often a chain of loaning of the word includes several languages. For example, Italian words *купол, кавалер, бензин, коридор* appeared in Russian language by being loaned from French, where they originally appeared from Italian.

In spite of the problems described there has never been any specialized computer system developed for automated etymological analysis.

There certainly are data mining systems that allow to reduce target area of the search, but the major part of the work is being done manually be people.

Considering the automation of the etymological analysis to be the primer goal the authors have developed fuzzy criteria for identification of the loaned words. These criteria are capable to discover words that have been loaned by using any of the following five ways: lexical-word-formative tracing; lexical-word-formative half-tracing; semantic tracing; transcription; transliteration.

Lexical-word-formation is defined in this paper as a literal translation of word parts (prefixes, root, suffixes) with exact imitation of a way of its formation and semantic. The words *кислород* and *водород* are an example that illustrates the usage of this method.

The result of criterion $\mu_f(w_{l_i}, w_{l_j})$ that defines the relationship extent between words $w_{l_i} \in l_i$ and $w_{l_j} \in l_j$ based on lexical-word-formation tracing is computed by the following algorithm:

1. Find all possible translations of the parts of the word w_{l_j} received as a result of the morphological analysis into language l_i;
2. Make fuzzy comparison of translations of the word w_{l_i} parts with the corresponding parts of the word w_{l_i};

3. If a number of translations (combinations of the translated parts of the word) for the parts of the word w_{l_i} forming a word w_{l_i} with average accuracy $\gamma > 60\%$ is found as a result of step 2 then $\mu_f(w_{l_i}, w_{l_j}) = \gamma$, otherwise $\mu_f(w_{l_i}, w_{l_j}) = 0$.

Lexical-word-formative half-tracing is carried out by literal translation of the foreign word parts and adding to them parts from analyzed language. For example, the word *гуманн-ость* in Russian language has been received by this method from a Latin root *human-us* and Russian suffix «*-ость*».

The result of criterion $\mu_h(w_{l_i}, w_{l_j})$ that defines the relationship extent between words $w_{l_i} \in l_i$ and $w_{l_j} \in l_j$ based on lexical-word-formation tracing is computed by the following algorithm:

1. Find all possible translations of the parts of the word w_{l_j} received as a result of the morphological analysis into language l_i;
2. Make fuzzy comparison of translations of the word w_{l_i} parts with the corresponding parts of the word w_{l_i};
3. If a number of translations (combinations of the translated parts of the word) for the parts of the word w_{l_i} forming a word w_{l_i} with a minimum accuracy more then 60% is found as a result of step 2 or a maximum accuracy of all found adequacies found as a result of step 2 is less then 60% , then $\mu_h(w_{l_i}, w_{l_j}) = 0$;
4. If only a part of the word w_{l_i} can be formed as a result of comparison taken on step 2 with an accuracy $\gamma > 60\%$ and the remaining part of the word can be formed by using grammars of the morphological analysis of language l_i then $\mu_h(w_{l_i}, w_{l_j}) = \gamma$, otherwise $\mu_h(w_{l_i}, w_{l_j}) = 0$.

Semantic tracing implies assignation of a new semantic meaning to a word from analyzed language under the influence of another language. For example, Russian word *картина* that designated "painting", "spectacle", under the influence of the English language began to be used in the meaning of "movie".

The result of criterion $\mu_s(w_{l_i}, w_{l_j})$ that defines the relationship extent between words $w_{l_i} \in l_i$ and $w_{l_j} \in l_j$ based on semantic tracing is computed by the following algorithm:

1. Enter all homonyms of words w_{l_i} and w_{l_j} into the sets $O_{w_{l_i}}$ and $O_{w_{l_j}}$ accordingly;
2. If $\left|O_{w_{l_i}}\right| = 1$ and/or $\left|O_{w_{l_j}}\right| = 1$, then $\mu_s(w_{l_i}, w_{l_j}) = 0$;
3. If $\left|O_{w_{l_i}}\right| > 1$ and $\left|O_{w_{l_j}}\right| > 1$, then for all $w_i \in O_{w_{l_i}}$:

   3.1 Make fuzzy comparison of the word's w_i synonyms and the synonyms of all words from the set $w_j \in O_{w_{l_j}}$;

   3.2. If there is at least one pair of the synonyms of words w_i and w_j that are congruent with an accuracy $\gamma > 60\%$, then $\mu_s(w_{l_i}, w_{l_j}) = \gamma$;

4. If there are no pairs of synonyms of the words w_i and w_j that are congruent with an accuracy $\gamma > 60\%$ discovered during the step 3, then $\mu_s(w_{l_i}, w_{l_j}) = 0$.

The result of criterion $\mu_\varphi(w_{l_i}, w_{l_j})$ that defines the relationship extent between words $w_{l_i} \in l_i$ and $w_{l_j} \in l_j$ based on transcription is computed by the following method:

If a congruency with an accuracy $\gamma > 60\%$ is a result of comparison of the $\varphi(w_j)$ and $\varphi(w_i)$, where $\varphi()$ is an operation of transcription then $\mu_\varphi(w_{l_i}, w_{l_j}) = \gamma$, otherwise $\mu_\varphi(w_{l_i}, w_{l_j}) = 0$.

The result of criterion $\mu_\tau(w_{l_i}, w_{l_j})$ that defines the relationship extent between words $w_{l_i} \in l_i$ and $w_{l_j} \in l_j$ based on transliteration is computed by the following method:

If a congruency with an accuracy $\gamma > 60\%$ is a result of comparison of the $\tau(w_{l_i}, l_j)$ and w_{l_j}, where $\tau()$ is an operation of transcription then $\mu_\tau(w_{l_i}, w_{l_j}) = \gamma$, otherwise $\mu_\tau(w_{l_i}, w_{l_j}) = \gamma$.

The final relationship extent between words w_{l_i} and w_{l_j} is defined as:

$$\lambda(w_{l_i}, w_{l_j}) = \max(\mu_f(w_{l_i}, w_{l_j}), \mu_h(w_{l_i}, w_{l_j}), \mu_s(w_{l_i}, w_{l_j}), \mu_\varphi(w_{l_i}, w_{l_j}), \mu_\tau(w_{l_i}, w_{l_j}))$$

The described criteria system covers all major ways of words loaning and allows the authors to define the model of the automated etymological analysis.

Authors' Information

Alla Zaboleeva-Zotova - 400066, CAD department, Volgograd State Technical University, pr. Lenina 28, Volgograd; e-mail: zabzot@vstu.ru

Ilya Prokhorov - 400066, CAD department, Volgograd State Technical University, pr. Lenina 28, Volgograd; e-mail: ilya.prokhorov@gmail.com