ON ZONAL MORPHOLOGICAL APPROACH TO NATURAL LANGUAGE TEXTS PROCESSING

D.V. Shlepakov, L.N. Shlepakov
Institute of Mathematics
National Academy of Sciences of Ukraine
Tereshchenkovska Str, 3
252601, Kyiv, Ukraine
Phone: +380 (44) 224 5563
Email: expertis@carrier.kiev.ua

ABSTRACT

We discuss some actual problems of natural language processing. We consider flexile language case. We propose zonal morphologically based model instead of traditional word based one. If we deal with flexile language there should be an interim layer of language units. We assume that morphs layer should be considered. It seems to be natural and non-restrictive. By theirs order in word we divide them on four categories or zones: prefixes \((P)\), roots \((R)\), suffices \((S)\), and endings \((E)\).

We introduce new term, namely semantic coverage. Semantic coverage is analogue of compact set over all possible words domain. Traditional single dictionary method assume that unique dictionary \(D_\Omega\) define semantic coverage \(\sigma^{(1)}(\Omega)\) of the language words domain. Let \(\sigma^{(4)}(\Omega) = \{D^{(Z)}_\Omega\}\) where \(Z\) means one of morphological zone: prefixes, roots, suffices, or endings. We denote \(Z = \{P | R | S | E\}\). We assume that coverage \(\sigma^{(1)}(\Omega)\) and coverage \(\sigma^{(4)}(\Omega)\) are linguistically identical, i.e. \(\sigma^{(4)}(\Omega)\) has only morphs those are part of some word in \(\sigma^{(1)}(\Omega)\). We consider compression rate of both methods in case of uniform compression. We obtain that compression rate of traditional single dictionary model less than proposed zonal one.

We survey some aspects of architecture of morphological processing system. We consider modified Huffman coding that used in facsimile hardware. We know that fax machine processes only black and white pixel series. Furthermore, they alternate all the time. We may map black and white pixel series to morphological zones. Huffman coding prefixes can be redefined to fit four zones structure. Another way to fit fax paradigm is two step appliance of Huffman coding, i.e. we join zones by pairs then apply the coding inside joint pairs, and finally, we use the coding for outside pairs.

We note that branching of multi-level system should be reasonable. All modifications of basic architecture should influence root part only. Variable part of root zone is regulated by threshold for overflow control. Other parts should be considered as unchangeable because of their constancy as morphological units of the language.

The perspectives and problems of flexile language texts processing are discussed too.