HP Laboratories Technical Report, 2013, N14

Dictionary and pattern-based recognition of organization names in Russian news texts

Solovyev V., Gareev R., Ivanov V., Serebryakov S., Vassilieva N.
Kazan Federal University, 420008, Kremlevskaya 18, Kazan, Russia

Abstract

This paper describes a part of the event extraction system which has been developed in collaboration with HP Labs Russia. The domain of input texts is business news feeds. One of the most important event participant types is 'Organization'. This paper is focused on the problem of organization names recognition in Russian news texts. Two approaches have been implemented. The first is dictionary-based. We propose an algorithm to make a dictionary from a set of legal body full names gathered from a government registry. The main problems with the dictionary matching are incorrect stemming and significant fraction of ambiguous names among dictionary entries. The second recognition approach is based on usage of local context clues and internal name words. These words constitute patterns which are intrinsic to organization names. These patterns enable recognition of non-dictionary names. We propose an algorithm to derive such patterns from the original dictionary. © 2013 Hewlett-Packard Development Company, L.P.

Keywords

Knowledge-based event extraction, Named entity recognition