Classification of Emotion Words in Russian and Romanian Languages

Marina Sokolova
Children's Hospital of Eastern Ontario
401 Smyth Rd., Ottawa, Ontario, Canada
msokolova@ehealthinformation.ca

Victoria Bobicev
Technical University of Moldova
Studentilor, 7, Chisinau, Moldova
vika@rol.md

Abstract

This paper presents a machine learning study of affective words in Russian and Romanian languages. We tag the word affective meaning by one of the WordNet Affect six labels *anger, disgust, fear, joy, sadness, surprise* and group into "positive" (*joy, surprise*) and "negative" (*anger, disgust, fear, sadness*) classes. We use the word spelling, a word form, to represent words in machine learning experiments to solve the multiclass classification and binary classification problems. The results show that the word form can be a reliable source of learning the affect.

Keywords: phonosemantics, sentiment analysis, machine learning

1 Motivation

Computational Natural Language Learning have been making steady progress in various aspects of Natural Language Processing (NLP). Many tasks have been successfully solved, e.g., document topic classification obtained accuracy comparable with human evaluation. However, some problems have been a challenge for algorithmic solutions, although humans routinely solve such tasks, e.g., spotting difference between terrible accident and terrific situation.

A fundamental, essential language characteristic is the word sense which is often recognized in a rather intuitive way. Senses of words given in machine-readable dictionaries sometimes are not adequate to what people have in mind. This inadequacy was demonstrated in the field of word sense disambiguation (WSD) where the machine- readable dictionaries failed to help in text understanding [5]. At the same time, some tools have become a success. WordNet[1], a public domain lexical knowledge base, is a powerful semantic network regularly used in word sense disambiguation. Another example is Roger's Thesaurus [2] which groups words together by implicit semantic relations. Such resources map word senses to certain explanations and connections with other words.

In the current work, we use machine learning algorithms to learn relations between word meanings and their sounds. A word as a *linguistic sign* can be attributed with two essential characteristics, the sound and meaning , where meaning refers to the word reference, i.e. the concept the word describes. For example, ball and its sound directly correlate with a round, soft object which is used to throw and catch around. Relations between the word sound and meaning are far from certain. In [3], the association between the word sound and its meaning is said to be arbitrary. In contrast, Phonosemantics, the theory of sound symbolism, is based on a hypothesis that relations exist between the two characteristics [13].

The goal of this work is to build lexical resources for Russian and Romanian languages based on the WordNet-Affect domains. The resources are then used to test the hypothesis that word form is relevant to meaning, in this case – the emotions the words convey. We build two data sets, Russian and Romanian respectively, based on the WordNet Affect emotion synsets [12]. To represent the data in machine learning experiments, we use the fact that in Russian and Romanian languages the word sounds directly correspond to the word orthography. Thus, we use the word spelling, a word form, as a substitution for its sound. Specifically, we use the letter form of *transliterated* Russian words and Romanian words for machine learning classification of words' affects. The word emotions are categorized into the WordNet Affect emotion classes *anger, disgust, fear, joy, sadness, surprise.* We solve multi-class and binary classification problems. to classify the words into the six classes and into binary (*joy, surprise* vs others) classes. We apply algorithms with different learning paradigms. The obtained empirical results show that, under certain conditions, the word form can be a reliable source of learning its affect.

Our study contributes to the development of much needed tools, as in recent years, most of the Internet use growth was supported by non-native English speakers. Starting in 2000, for non-English speaking regions, the growth has surpassed 3,000 % compared with the over-all growth of 342%.[3] Consequently, the amount of text data written in languages other than English rapidly increased. This surge has prompted the demand for automated text analysis. The tool development progressed for some languages (French, German, Japanese, Chinese, Arabic), whereas some languages (Eastern European), have not yet attracted much attention from the NLP and Text Data Mining community. The presented study contributes to filling the gap.

[1] http://wordnet.princeton.edu/

[2] http://thesaurus.reference.com/

[3] http://www.internetworldstats.com/stats.htm