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 multi-class 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

2 Phonosemantics

Words (rabbits) and morphemes (rabbit,-s) are commonly accepted as the meaning-bearing units which provide association between sound and meaning [7]. However, a hypothesis that smaller units, phonemes and phonetic features, can bear meaning has found supporters in Phonosemantics research [6, 8, 9]. Based on the notion of *distinctions* [15], three types of sound meaning have been suggested [8, 9]:

Onomatopoeia is the imitation of a sound like, for example, in roar or moo.

Clustering is an effect of the semantic association.

True Iconism is the visceral effect of the sound on a person.

Semantic properties of the phonetic features are unconsciously learned by a child, e.g. the smallness implicit in the /i/ sound and the wetness implicit in the /w/[4]. Once the word is assigned to its referent, i.e., concept, this intuitively apprehended semantics is masked by the referent, but it does not cease to act altogether. The effect of the sound is still influencing the word. This influence remains on the unconscious level therefore it is difficult to pick out phoneme senses.

In this work, we hypothesize that a word's form and sound have certain relations with its meaning. Consequently, the meaning of a word is in part inherited from its form. For example, slide is a smooth motion, the smoothness and slipperiness so common in /sl/ shows up in the actual referent for slide [9]. In words with a more specific reference, the component of the reference is more salient; consequently, the sound-meaning part is less salient. For example, words which denote material objects (house, train) have senses dominated by the referents. On the other hand, words which denote abstract concepts as sensations, feelings and emotions keep more sound semantics in their forms (anger, joy, agitation). Thus, words describing similar sentiments should have something similar in their sounding (vex, worry), whereas words representing opposite feelings should have much less in common (disgust, elation).

Relations between the form and meaning of English emotional words were analyzed in [10]. The authors applied K-Nearest Neighbor, a prototype-based learner, to classify affective words into multi-class and binary emotion categories. The empirical results showed that the word forms for English words expressing the same emotion are alike in certain ways. Our current study differs from [10] as follows:

1. We study Russian and Romanian emotional words. Both languages are Eastern European, belonging to *Slavic* and *Latin* families, respectively.

2. We analyze the learning abilities of different paradigms, i.e. probability-, prototype-, decision- and optimization-based algorithms.

[4] http://www.trismegistos.com/MagicalLetterPage/

Table 1: *Translation of the WordNet 05573914 n*

English	Romanian
preference	preferinta
penchant	inclinatie, slabiciune
predilection	predilectie
taste	a avea gust, a gusta, a cunoaste; a gusta; a degusta (un aliment), degustare, ...

3 Lexical Resources

WordNet Affect WordNet-Affect [5] is a lexical resource which is based on the lexical knowledge of the (English) WordNet. WordNet-Affect contains words which convey affects. A number of affective labels (*a-labels*) were manually assigned to the synonym sets (*synsets*) of nouns, adjectives, verbs, and adverbs. The words with the Emotion tag were fine-grain annotated using six labels: *joy, fear, anger, sadness, disgust, surprise* [12]. The six emotion tags were adapted from the study of human non-verbally expressed emotions [4]. We used the WordNet Affect data provided at the SemEval-2007 "Affective Text" [11].

Russian and Romanian WordNet Affects We translated the WordNet Affect synsets into Russian and Romanian. We applied a three-step approach:

Translation We manually translated every word in the six WordNet Affect emotion categories; Table 1 gives an example of a synset translation. We omitted word combinations (get happy), collocations and idioms. The other restriction was that the translations were related to the emotion of the synset. We postponed part-of-speech correspondence till the later phases. For the Romanian data set, we used the on-line dictionary Dexonline [6] to obtain all synonyms of the translated words.

Building the word sets to form the word sets for analysis, we made a list of all the translations. We edited them to delete words which meanings were not close to the emotion, e.g., for taste, only preferinta was left, all the food references were removed (Table 1). We removed duplicate translations as well. As a result, we built six sets of Russian words and six sets of Romanian words expressing the WordNet Affect emotions.

Reducing the number of the paronymous words Russian and Romanian languages are rich in derivations (schastlivyi, schastliven'kii); there were sometimes four, five – or more – words with the same root. We removed all the paronymous words. Note that Romanian and Russian languages allow letter alternation in the word root. Thus, we kept two words per root (zlo, zliti) if the number of matching letters was < 3.

[5] http://wndomains.itc.it.
[6] http://dexoline.ro

Table 2: *Data sets of affective words: English, Russian, Romanian.*

Classes	English data				Russian data			Romanian data		
	# synsets	% synsets	# words	% words	# words initial	# words	% words	# words initial	# words	% words
anger	128	21.0	318	20.7	149	105	13.0	316	151	25.0
disgust	20	3.3	72	4.7	46	31	5.0	93	43	3.9
fear	83	13.5	208	13.5	118	71	14.6	123	55	12.8
joy	228	37.2	539	35.1	253	183	36.2	510	211	37.7
sadness	29	4.7	309	20.1	217	128	25.6	241	111	16.2
surprise	124	20.3	90	5.9	54	29	5.6	91	48	4.4
Total	612	100.0	1536	100.0	837	547	100.0	1374	619	100.0

Table 3: *Distribution of the affective nouns in the Russian and Romanian data.*

Russian Data			Romanian Data		
Classes	# nouns	%	Classes	# nouns	%
anger	39	13.0	anger	51	25.0
disgust	15	5.0	disgust	8	3.9
fear	44	14.6	fear	26	12.8
joy	109	36.2	joy	77	37.7
sadness	77	25.6	sadness	33	16.2
surprise	17	5.6	surprise	9	4.4
Total	301	100.0	Total	204	100.0

Table 2 describes the WordNet-Affect synsets used in our work: **# synsets** presents the initial number of the English synsets, **% synsets** shows per cent for each class, **# words** – the unique words count for the English, **% words** – per cent of words for each emotion, **# words initial** presents the number of the Russian words before the removal of paronymous words, **# words** and **% words** list counts and per cent of the Russian words which were used in classification experiments, **# words initial, # words, % words** list the similar information for the Romanian data set. The data sets are available for research purposes. [7]

Further, the Russian and Romanian sets were *each* split into *nouns* and the other Part-of-speech. The experiments were conducted on *nouns* only; see Table 3 for details. Other part-of-speech are left for future analysis.

Previous Work Romanian WordNet was created during BalkaNet [14], a multilingual database comprising of the individual WordNets for the Balkan languages. It assigns synsets with three sentiment scores (positive, negative, objective).[8] For Russian resources, little information is available. RussNet [1] and Russian WordNet[2] are non-commercial projects. Two commercial projects are RuThes[9], an informational thesaurus, and the Russian WordNet Novosoft[10].

4 Empirical Results

We defined two *supervised* problems: (i) to classify a word as "positive" (*joy, surprise*) or "negative"(*anger, disgust, fear, sadness*) ; (ii) to classify a word with one of the six affect labels. We constructed four labelled data sets: (i) the *transliterated* Russian words (the six classes); (ii) the *transliterated* Russian words (the two classes);(iii) the Romanian words (the six classes); (iv) the Romanian words (the two classes). For each data set, we built seven representations. Five representation omit the word letter order: **Letters-All**, every letter that appeared in the word had its occurrence counted; **Vowels**, only vowels that appeared in the word were counted; **Consonants**, only consonants that appeared in the word were counted; **Letters-3**, words were represented by occurrences of the first three letters; and **Letters-4**, words were represented by occurrences of the first four letters. Two representations use the word letter order: **OrderLetters-3**, words were represented by the occurrences of the first three letters and their order; and **OrderLetters- 4**, words were represented by the occurrences of the first four letters and their order.

We applied the following algorithms: probability-based (Naive Bayes, Bayes Nets), prototype-based (k-Nearest Neighbor), decision-based (C4.5 (decision tree) and PART(decision list)), and optimization (Support Vector Machines).[11] For binary classification, we report *Accuracy*, *Precision*, *Recall* and the balanced *Fscore*. For multi-class classification, we report the *macro*-average *Precision*(P), *Recall*(R), and the balanced *Fscore*(F). To avoid the bias towards the majority class (*joy*), we report *Accuracy* obtained with the highest *Fscore*. Tables 4 and 5 list the best results of SVM, KNN, C4.5, PART. Naive Bayes and Bayes Nets performed considerably poorer. Both tables omit their results. SVM performed more accurately than the other learners. Only on multi-classifying the Russian words, SVM was outperformed by KNN.

The learning results differ for the two languages.The Russian words were classified more accurately when their identification was more precise: the overall best *Accuracy* corresponds to the overall best *Fscore*. The Romanian emotion words can be accurately classified without the highest precision: the overall best *Accuracy* and the overall best *Fscore* are obtained by different classifiers. The Russian words were classified the best on the first three letters, without indicating the letter order. The Romanian words were better classified if represented by vowels (the multi-class tie), the ordered first four letters (binary, the multi-class tie); the highes precision was obtained on consonants (binary) and all the letters (multi-class).

[7] http://lilu.fcim.utm.md
[8] http://sentiwordnet.isti.cnr.it/
[9] http://www.cir.ru
[10] http://research-and-development.novosoft-us.com
[11] the Weka software: http://www.cs.waikato.ac.nz/ml/weka/

Table 4: *Binary classification of the affective words, in per cent. Table reports the best Accuracy and corresponding Fscore measures for each algorithm. The overall best Accuracy and Fscore for the data are in* **bold**. *Baseline Accuracy for the Russian data – 58.7 %, for the Romanian data – 57.8 %.*

Russian Data

Feature Sets	Algorithms											
	SVM				KNN				Decision-based			
	Acc	F	Pr	R	*Acc*	F	Pr	R	*Acc*	F	Pr	R
Letters-All	62.6	72.2	64.1	82.7	61.3	72.8	62.0	88.3	65.2	70.7	69.9	71.5
Vowels	62.0	71.8	63.5	82.7	63.2	72.4	64.8	82.1	60.7	69.4	63.8	76.0
Consonants	63.0	72.1	64.6	81.6	62.2	74.2	62.4	91.6	58.0	67.0	62.2	72.6
Letters-3	**71.0**	**78.9**	68.5	93.2	70.0	77.6	68.7	89.3	64.4	71.4	67.2	76.3
Letters-4	67.7	74.9	68.1	83.3	66.3	73.8	67.3	81.6	64.0	71.1	67.0	75.7
OrderLetter-3	66.6	76.7	64.9	93.9	66.6	76.7	64.9	93.9	62.3	70.6	65.1	77.1
OrderLetter-4	67.2	77.2	65.3	94.4	64.3	75.8	62.9	95.5	63.3	69.9	67.4	72.6

Romanian Data

Feature Sets	Algorithms											
	SVM				KNN				Decision-based			
	Acc	F	Pr	R	*Acc*	F	Pr	R	*Acc*	F	Pr	R
Letters-All	64.7	70.0	68.9	71.2	65.7	72.2	67.9	77.1	60.8	66.7	65.6	67.8
Vowels	60.8	68.8	63.8	74.6	61.8	68.8	65.2	72.9	57.3	66.9	60.7	74.6
Consonants	64.7	**73.9**	64.6	86.4	59.8	67.7	63.2	72.9	60.3	65.8	65.5	66.1
Letters-3	59.8	69.6	61.8	79.7	58.3	68.4	60.9	78.0	57.8	65.6	62.1	69.5
Letters-4	61.3	72.7	61.4	89.0	62.8	71.4	64.2	80.5	58.3	66.9	61.9	72.9
OrderLetters-3	63.7	72.6	64.5	83.1	61.8	70.0	64.1	77.1	62.3	68.8	65.9	72.0
OrderLetters-4	**67.8**	73.0	70.6	75.4	64.2	73.3	64.5	84.7	62.8	68.6	66.9	70.3

Table 5: *Multi-class classification of the affective words, in per cent. Table reports the best Accuracy and corresponding* macro-*average Fscore measures for each algorithm. The overall best Accuracy and Fscore for the data is in* **bold**. *Baseline for the Russian data: Fscore – 8.9 %, Precision– 6.0 %; baseline for the Romanian data:Fscore– 9.7 %, Precision–6.2 %.*

Russian Data

Feature Sets	Algorithms											
	SVM				KNN				Decision-based			
	Acc	F	Pr	R	*Acc*	F	Pr	R	*Acc*	F	Pr	R
Letters-All	37.4	16.6	17.2	20.4	33.4	21.5	26.4	21.5	32.1	21.0	22.5	20.9
Vowels	35.4	13.3	11.1	17.9	31.8	16.0	15.9	17.6	28.5	16.3	16.8	17.3
Consonants	38.9	15.1	12.9	19.8	36.4	15.9	22.0	19.1	27.2	16.7	18.5	17.2
Letters-3	40.0	27.1	32.6	27.4	**40.5**	**29.3**	35.5	28.3	38.9	25.8	32.8	25.4
Letters-4	38.7	16.7	18.0	20.1	35.7	19.0	18.4	20.6	37.9	19.6	21.4	21.8
OrderLetter-3	39.3	22.4	27.5	24.2	38.0	28.6	31.4	27.8	36.4	26.2	29.4	25.8
OrderLetter-4	35.4	15.8	19.6	19.0	36.4	22.6	23.5	23.4	32.8	19.7	22.7	20.7

Romanian Data

Feature Sets	Algorithms											
	SVM				KNN				Decision-based			
	Acc	F	Pr	R	*Acc*	F	Pr	R	*Acc*	F	Pr	R
Letters-All	35.8	20.1	20.6	20.8	34.8	19.2	23.8	19.8	38.2	**23.8**	23.7	24.5
Vowels	**40.7**	18.8	22.7	21.0	37.3	20.9	22.0	21.6	39.2	19.9	20.4	21.3
Consonants	35.9	20.1	20.7	21.0	29.4	18.6	21.8	20.3	36.3	21.8	21.4	22.4
Letters-3	38.2	17.3	16.3	20.4	37.8	20.4	23.0	21.6	40.2	17.1	20.0	20.2
Letters-4	37.3	18.7	18.7	19.9	34.3	17.8	19.3	19.1	35.8	19.2	18.9	20.2
OrderLetters-3	36.8	19.7	19.9	20.7	36.8	19.2	19.2	20.8	36.3	19.1	18.4	20.4
OrderLetters-4	**40.7**	21.4	21.0	22.9	38.2	19.5	22.0	21.1	36.3	20.0	20.1	20.0

5 Discussion and Future Work

We have presented a study of the relations between word form and meaning for affective words. We have studied emotion words in Russian and Romanian. The obtained empirical results show the reliability of our learning approach. On the Russian and Romanian data sets, the applied algorithms performed *considerably better* than baselines. Although the difference in data sets does not allow a direct comparison, our results appear to be more accurate and precise than the results for the English affective words [10].

Based on the results of this study, we propose that there is similarity among the forms of words that express the same emotion: the word form similarity was captured by machine learning algorithms which classified words according to their emotion tags. We also sought a better word form presentation. In Russian and Romanian languages, word spelling can be considered as a word phonetic equivalent. This feature allowed us to limit the search to letter-based representations. It should be noted that letter representations provided better results for English affective words [10], although in English correspondence between the letters and phonemes is not unique, i.e. the same letter can represent different sounds depending on the neighboring letters (cat – [k],certain–[s]). Thus, we can conclude that for phonosemantic classification, letter representations may provide relevant information about the word form.

For future studies, we plan to concentrate on features which better discriminate among emotion classes. We also want to determine which sounds better correlate with the conveyed emotions. Our current hypothesis is that for every emotion there are several classes of words that share common phonological features. For example, the sound z is present in Russian words with meaning of amazement; the *transliterated* sound sh can be found in Russian words representing a kind of stupefaction (there is no absolutely precise translation of these words in English). Note that the exact translation of the English word stupefy is ostolbenet'. Hence, the *transliterated* word and its translation share the combination of sounds st. These are preliminary remarks. A thorough analysis will be able to demonstrate the existence – or the absence – of semantic relations between words with common phonetic features. Another venue would be to expand our current study to part-of-speech other than nouns. We also are interested in conducting *human evaluation.*, i.e., based on the listed word representations, query native speakers about evoked emotions.

6 Conclusions

We have constructed Russian and Romanian word sets based on the WordNet Affect domains. Although multiple efforts have been made to create lexical resources similar with English WordNet for other languages[12], lexical resources for Eastern European languages are still limited. Our study contributes to the development of the resources.

We have shown that the word forms of *transliterated* Russian words and Romanian words allow for a reliable classification of their emotions, in both multiclass and binary settings. The empirical results support our hypothesis that the word spelling is relevant to the emotion that the word conveys. The obtained results can further be used in the nested learning of sentiments in Russian and Romanian texts.

[12] http://multiwordnet.itc.it/english/home.php

Acknowledgements

This work was in part supported by the Natural Sciences and Engineering Research Council of Canada and the RANLP travel grant. We thank Elizabeth Jonker for helpful comments on the text. We thank Victoria Maxim and Natalia Burciu for assistance with building the Romanian and Russian data sets.

References

[1] I. Azarova, O. Mitrofanova, A. Sinopalnikova, M. Yavorskaya, and I. Oparin. Russnet: Building a lexical database for the russian language. In *Proceedings of the Workshop on Wordnet Structures and Standardisation and How this affect Wordnet Applications and Evaluation*, 2002.

[2] V. Balkova, A. Sukhonogov, and S. Yablonsky. Russian wordnet. In *Proceedings of the Second Global Wordnet Conference*, 2004.

[3] F. de Saussure. *Cours de linguistique générale*. Harrassowitz, Wiesbaden, 1916.

[4] P. Ekman. An argument for basic emotions. *Cognition and Emotion*, 6:169–200, 1992.

[5] N. Ide and J. Veronis. Word sense disambiguation: The state of the art. *Computational Linguistics*, 24(1):1–40, 1998.

[6] R. Jakobson and L. Waugh. *The Sound Shape of Language*. Indiana University Press, 1979.

[7] D. S. Jurafsky and J. H. Martin. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Prentice Hall, New Jersey, 2000.

[8] M. Magnus. *Gods of the Word : Archetypes in the Consonants*. Truman State University Press, 1999.

[9] M. Magnus. *What's in a Word? Studies in Phonosemantics*. PhD thesis, University of Trondheims, Norway, 2001.

[10] V. Nastase, M. Sokolova, and J. Sayyad Shirabad. Do happy words sound happy? a study of the relation between form and meaning for english words expressing emotions. In *Proceedings of the Recent Advances in Natural Language Processing (RANLP'2007)*, pages 406 – 410, 2007.

[11] C. Strapparava and R. Mihalcea. Semeval-2007 task 14: Affective text. In *Proceedings of the 2008 ACM symposium on Applied computing*, 2008.

[12] C. Strapparava, A. Valitutti, and O. Stock. The affective weight of the lexicon. In *Proceedings of the 5th International Conference on Language Resources and Evaluation*, pages 474–481, 2006.

[13] R. Tarte and M. O'Boyle. Semantic judgements of compressed monosyllables: Evidence for phonetic symbolism. *Journal of Psycholingual Research*, 11(3):183–196, 1982.

[14] D. Tufis, B. Mititelu, L. Bozianu, and C. Mihaila. Romanian wordnet: New developments and applications. In *Proceedings of the 3rd Conference of the Global WordNet Association*, pages 337–344, 2006.

[15] W. von Humboldt. *Über die Verschiedenheit des Menschlichen Sprachbauesund ihren Einflu auf die geistige Entwicklung des Menschengeschlechts*. Druckerei der Könglichen Akademie, 1836; reprinted: 1960.