Extraction of High-Level Semantically Rich Features from Natural Language Text*

Dasha Bogdanova

University of St. Petersburg
dasha.bogdanova@gmail.com

Abstract. To represent a text, Natural Language Processing applications are to determine and extract from the text features that are essential for the particular task. Although high-level features seem to be promising for many tasks, they were rarely addressed, since the extraction of those features is a big challenge. This thesis aims at extracting high-level semantically rich features from natural language text. The algorithms we will propose will enable development of novel applications in different areas.

1 Introduction

In many NLP tasks documents are represented as feature vectors. These vectors can then serve as an input to various algorithms such as e.g. document clustering and classification. These features are to reflect essential for the particular task characterics of the documents. For example, the topic of a document could hardly be reflected by the average sentence length. Though sentence length could be utilized in authorship analysis, since some authors are known for using very long sentences (e.g. L. Tolstoy) while others prefer shorter ones (e.g. E. Hemingway).

The most widely used features are primarily lexical and character ones, those that consider a text as a sequence of words and characters respectively. Namely, word frequencies [7, 10], vocabulary richness [10], n-grams [12], letter frequencies [6], character n-grams [24], etc. A very big advantage of those low-level features is that they are easy to extract automatically. High-level features capturing not only the symbolic information but more semantics of a text often appear to be more promising while solving different tasks, but modern NLP tools do not provide accurate extraction of those features. Therefore they were very rarely exploited.

The goal of the thesis is to develop a number of algorithms for high-level semantically rich features extraction and to evaluate these features in terms of their applicability to different NLP tasks.

* This work is partially supported by Russian Foundation for Basic Research under grant 10-07-00156 and Google Research Award

2 The State of the Art

2.1 Figurative Language Extraction

One of the first attempts to automatically process figurative language is presented in [8]. The described system recognizes metaphor and metonymy based on selectional preferences. For example, the subject of *drink* in its literal meaning should be animate. It is not satisfied in case of *My car drinks gasoline*, where the verb appears as a metaphor. The method requires manually annotated corpora and therefore it is difficult to extend.

The TroFi system presented in [4] aims at identifying figuratively used verbs. The authors construct two seed sets, one of them contains literal usages, and the other contains metaphorical ones. The algorithm measures the similarity between the context of the utterance in question and the seed sets and labels the utterance as literal or metaphorical depending on what set is closer.

The method proposed in [9] distinguishes between literal and metaphorical usages. The approach is based on Maximum Entropy classification. The training data consists of manually annotated instances of MOTION and HEALTH verbs from the Wall Street Journal corpus.

Another system for automatic metaphor identification is described in [22]. Starting with a small seed set of annotated metaphorical expressions, the system is able to identify metaphors in a large corpus by applying verb and noun clustering.

In [5] we have presented an approach to detect figurative language in general. Given an expression in a context, the proposed algorithm is to decide whether the expression is used metaphorically or literally. The underlying idea of the algorithm is as follows: if there is a significant difference between usual sense of an expression and a sense of its context, the expression is likely to be used figuratively. We are going to extend this study and to propose a number of methods for extracting different types of figurative language. Unlike the majority of the previous work on the subject, we do not focus only on verbs. We will consider different parts of speech as well as metaphorically used multi-word expressions.

2.2 Sentiment Extraction

Sentiment analysis is a broad research area with various applications such as, for example, product and movie review mining. The problems of identifying opinionated documents and detecting their polarity have been actively addressed during the last years [26, 17, 27]. The problem of fine-grained emotion annotation was defined at the SemEval 2007 task on "Affective Text" [25]: given a set of news titles, the system is to label each title with the appropriate emotion out the following list: ANGER, DISGUST, FEAR, JOY, SADNESS, SURPRISE. The results obtained by participating systems show that the problem is very difficult and requires future study.

In contrast to the studies mentioned above, we consider sentiment as a high-level feature that can be exploited by other applications. In these terms, the most related to ours is the study described in [18]. The method is based on the idea that opinions of different authors are expressed differently. The authors show that opinion mining techniques can be improved by considering authors separately.

In this thesis, we also refer to the idea of jointly using authorship analysis and sentiment detection. The study [18] proves that opinion mining can benefit from authorship attribution techniques. We believe that the extraction of sentiments could as well improve authorship analysis, since emotions influence writing style.

2.3 Extraction of Psychological Characteristics

The problem of detecting personal psychological characteristics by analysing writing style has been rarely addressed. A study of the problem in question is presented in [1]. It shows that psychological characteristics of a person are reflected in his or her writing style. The authors used student essays as an experimental data, all students also had to fill in a questionnaire as to determine their "Big Five" personality dimensions: neuroticism, extraversion, openness, conscientiousness and agreeableness [11]. The results on neuroticism detection are presented and a relatively high accuracy rate of 65.7% is achieved.

In this thesis, we plan to consider these characteristics as well as other, such as e.g. learning style.

3 The Proposed Approach

The topic of this research is very broad and could not be completely covered in the thesis, therefore we focus our attention only on some particular problems and approaches within the topic. The thesis is to propose a number of algorithms for features extraction, that as we believe will enable development of various novel algorithms based on high-level features, such as e.g. authorship analysis, genre identification and opinion mining.

3.1 The Features

We describe below the features we plan to study.

Figures of Speech Figurative language refers to words that deviate from their literal meaning. Figurative language can be represented by metaphor, metonymy, simile, irony, etc. In this thesis we plan to address the extraction of the following figures of speech:

- *Metaphor.* It is a figure of speech that arises out of an analogy between two domains or ideas, thus a concept is described in terms of another concept's domain. The examples of metaphor: “She is a *sunshine*”, “My heart is *dancing*”.

– *Metonymy.* Similarly to metaphor, metonymy substitutes a name of a concept with another one. Although metaphorical substitution is based on similarity, whereas in metonymy, it is based on association or relatedness. The typical metonymic expressions include using the name of a place instead of some people or events associated with the place: "*Russia* decided to invest in roads", "In the time of *Vietnam*".
– *Irony.* It appears when a literal meaning is opposite to what is actually meant. For example, saying "This is *great* we have to wait another day", while a speaker is not really happy about the fact.

It has been shown by previous research that figures of speech are a frequent phenomenon of a language [9]. Being able to capture figures of speech, is essential for many applications, such as e.g. machine translation and dialogue systems. Moreover, the usage of figurative language can pertain to the author's writing style, therefore this type of features can be utilized by authorship analysis applications. The study described in [19] shows that figurative language extraction can be successfully applied to the tasks of sentiment analysis, since metaphorical expressions tend to be opinionated. Furthermore, this area will benefit from irony detection, because ironical utterances of a word have the opposite polarity from literal meaning, e.g. positive adjective *good* is negative while used ironically.

Sentiments in Text Sentiment analysis is a broad area which deals with computational processing of sentiment and opinion. It has various applications from political opinion mining to analyzing product reviews. Aside from the fact that sentiments in text are studied by an independent research area, we believe they can serve as high-level features for various other applications. Stylistic tasks such as authorship analysis can benefit from exploiting sentiment-based features, since different people tend to express their emotions differently and this fact can be utilized to distinguish between their writing styles.

Psychological Markers Given a text, we are to determine psychological characteristics of an author of the text. The results reported in [1] show that different psychological properties of a person can be learned from his or her writing style. The authors considered the "Big Five" personality dimensions in their study. In addition to this personal properties, we are going to consider learning style. In psychological literature, people are usually divided into visual, audial and kinesthetic learners [3]. Visual and audial learners learn primarily by seeing and hearing things respectively, while kinesthetic learners are those who learn best by feeling and doing.

This type of features can be utilized in psychology as well as in authorship profiling as described in [1].

Potential applications of the algorithms include authorship analysis, opinion mining.

4 Experiments

So far, we have performed only a small piece of the experiments, concerning gender identification and learning style detection. The experiments are desribed below.

4.1 Gender Detection

Men and women often express their emotions differently [16]. We expect this fact should have an impact on the writing style of different genders. Therefore, we hypothesize that opinionated lexicon can be used to distinguish between genders. In our experiments we both considered opinionated lexicon as the only feature and while combined with features proposed by previous research.

We used dataset presented in [21], it is available at the website of one of the authors. First, we selected only those blogs which have between 10 and 30 thousands of words. This set contained 1138 female-authored blogs and 1125 authored by males. Then we selected only words opinionated according to SentiWordnet [2] and applied Naive Bayes classification to the data. To estimate the accuracy, we used 10-fold cross validation. The results are presented in Table 1.

Table 1. Detailed Accuracy By Class for classification using all opinionated words

Class	TP Rate	FP Rate	Precision	Recall	F-measure
male	0.53	0.27	0.66	0.53	0.59
female	0.73	0.47	0.61	0.73	0.67

As it was defined in [14], we then calculated *gender score* of each feature. Feature weight is the probability of seeing this feature in a given category (gender).

$$Weight_C(F) = P(F|C) \approx \frac{Count(F)}{Count(C)}$$

$$GenderScore(F) = \frac{Weight_{female}(F)}{Weight_{female}(F) + Weight_{male}(F)}$$

Thus, if the gender score of a feature is closer to 1, the feature is more representative for female gender. And if the value is closer to 0, the feature is more representative for males. The most discriminative features (features with the highest and the lowest gender scores out of the features with total frequency more than 600) are shown in Table 2.

We then selected only those features that are more representative either for males or for females, i.e. those which have gender score closer to 0 or 1. As it is shown in Table 3, the highest F-measure is achieved when we exclude features with gender scores falling into the following intervals: (0.2; 0.7) and (0.2; 0.8). The confusion matrix and detailed accuracy for the latter case are presented in Table 4 and Table 5 respectively.

Table 2. The most representative opinionated words for each gender

	word	gender score
female	fabulous	0.76
	cute	0.73
	cry	0.72
male	comic	0.29
	liberal	0.27
	victory	0.27

Table 3. F-measure of the gender classification using opinionated words with gender score out of the specified interval.

Endpoints of the excluded interval	*0.1*	*0.2*	*0.3*	*0.4*
0.6	0.71	0.74	0.72	0.68
0.7	0.74	**0.75**	0.72	0.67
0.8	0.74	**0.75**	0.71	0.67
0.9	0.74	0.74	0.70	0.67

The presented results show that opinionated lexicon can be a remarkable feature for gender detection, though it is not reliable enough to serve as the only feature of a classifier. Therefore, we decided to combine opinionated lexicon with other features. In [21] a list of 30 words was suggested to distinguish between genders. The accuracy of the gender classification using this list is presented in Table 6.

Table 4. Confusion matrix for the gender classification using opinionated words with gender score $\notin$ (0.2; 0.8)

	true male	true female
classified as male	711	131
classified as female	414	1007

We have also conducted experiments using the combined feature list, it included both the opinionated words and the words from the list suggested in [21]. The F-measure values are shown in Table 7. This time the highest F-measure was achieved when the features with gender score falling into (0.2; 0.9) were excluded, though the improvement over classification based only on opinionated lexicon is not sufficient.

4.2 Learning Style Detection

In psychological literature, three main types of learners are defined: visual, audial and tactile (kinesthetic) learners according to the type of information they perceive better [3].

Table 5. Detailed Accuracy By Class, gender score $\notin$ (0.2; 0.8)

Class	TP Rate	FP Rate	Precision	Recall	F-measure
male	0.63	0.12	0.84	0.63	0.72
female	0.88	0.37	0.71	0.88	0.79

Table 6. Detailed Accuracy By Class for the gender classificating using the list of words presented in [21]

Class	TP Rate	FP Rate	Precision	Recall	F-measure
male	0.63	0.19	0.76	0.63	0.69
female	0.81	0.37	0.69	0.81	0.74

We hypothesize that the learning style influences the writing style of a person. In order to test the idea we used the same collection of blog posts as in the previous experiment. For each learning style we have defined a list of marker words. The naive way to construct such lists is to obtain hyponyms of the corresponding categories: *visual property* and *visual perception*, *tactile perception*, *taste sensation* and *taste property*, *auditory sensation*, *sound* and *sound property*. Thus, the list of visual markers contained *light*, *darkness*, colors such as *pink* or *violet*, etc., with the total number of 370. The audial list contained *whistle*, *clop* and *bark* as well as other sounds and such words as *music*, *chorus* and so on, with the total number of 354. The list of tactile markers included *sweet*, *sour*, *relish*, etc., the total number of the tactile markers is 123. We stemmed the words in the obtained lists and estimated the number of those words in given blog posts. The numbers of each category markers in the blog posts, normalized over the total amount of words in a post and the number of markers in the category, are presented in Figure 2.

The obtained results correspond to the ideas described in previous psychological research [15, 23]. According to [15, 23] a little more than one half of people does not have one prefered learning style, while others demonstrate prevalence of the only style. Among those, about 80% are kinesthetic learners.

Since this topic has been rarely addressed in the past, there is no avalaible benchmark for the task. Therefore complete in vitro evaluation is impossible

Table 7. F-measure of the gender classification using opinionated and frequent words with gender score out of the specified interval.

Endpoints of the excluded interval	*0.1*	*0.2*	*0.3*	*0.4*
0.6	0.72	0.74	0.72	0.68
0.7	0.73	0.75	0.73	0.68
0.8	0.74	0.75	0.72	0.68
0.9	0.75	**0.76**	0.72	0.67

on this stage. In this thesis, we are going to provide in vivo evaluation of the learning style detection in the authorship analysis tasks.

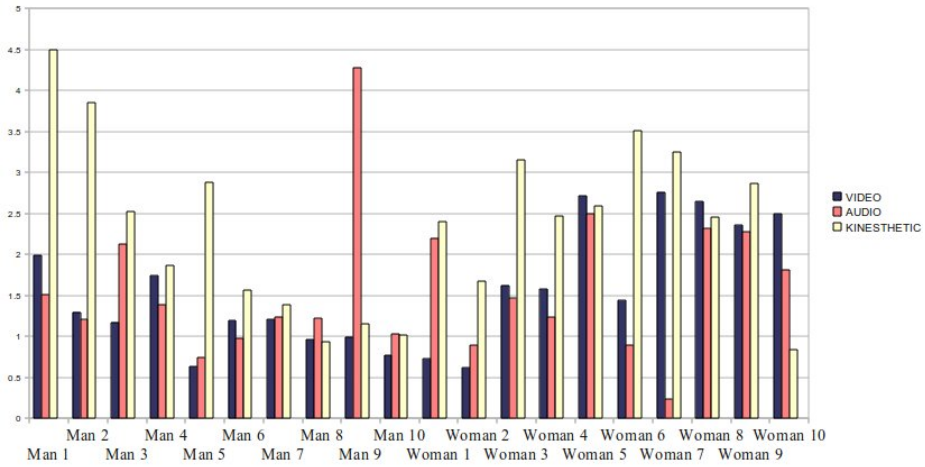

Fig. 1. The proportion of each learning style markers in blog posts

5 Conclusions and Future Work

The thesis we are working on will propose a number of algorithms for the extraction of high-level semantically rich features. Such features are known to be promising for various tasks including machine translation, question answering, authorship attribution, opinion mining, etc. We believe these algorithms will enable development of novel approaches different tasks. In order to test this assumption, we plan to evaluate the features in question in terms of their helpfulness while solving authorship analysis and opinion mining tasks. Thus, we will continue the work on the thesis in the following directions:

- **Feature Extraction**. We plan to propose several algorithms for feature extraction and to provide in vitro evaluation of these algorithms if possible.
- **Feature Integration**. As to provide in vivo evaluation of the feature extraction algorithms, we plan to focus on Authorship Analysis and Opinion Mining

In the next half a year we plan to address the problem of detecting author's gender. In this paper, we have considered opinionated lexicon as a feature to predict author's gender. According to the conducted experiments, authors of different genders tend to use different opinionated words, though the accuracy of the classifier based only on these features is less than the accuracy of state-of-art

methods. Combination of opinionated lexicon with the feature words described in [21] gave only insufficient improvement. Thus, we plan to consider other features. In particular, slang is described in [20] as reliable feature to predict author's gender.

References

1. S. Argamon, M. Koppel, J. W. Pennebaker, and J. Schler. Automatically profiling the author of an anonymous text. *Commun. ACM*, 52:119–123, February 2009.
2. S. Baccianella, A. Esuli, and F. Sebastiani. Sentiwordnet 3.0 : An enhanced lexical resource for sentiment analysis and opinion mining. *Proceedings of the 7th Conference on Language Resources and Evaluation LREC10*, pages 2200–2204, 2008.
3. J. S. Beaudry and A. Klavas. Survey of research on learning styles. *Educational Leadership*, 2(2):75–98, 2002.
4. J. Birke and A. Sarkar. A clustering approach for nearly unsupervised recognition of nonliteral language. In *In Proceedings of EACL-06*, pages 329–336, 2006.
5. D. Bogdanova. A framework for figurative language detection based on sense differentiation. In *Proceedings of the ACL 2010 Student Research Workshop*, ACLstudent '10, pages 67–72, Morristown, NJ, USA, 2010. Association for Computational Linguistics.
6. O. de Vel, A. Anderson, M. Corney, and G. Mohay. Mining e-mail content for author identification forensics. *SIGMOD RECORD*, 30:55–64, 2001.
7. J. Diederich, J. Kindermann, E. Leopold, G. Paass, G. F. Informationstechnik, and D.-S. Augustin. Authorship attribution with support vector machines. *Applied Intelligence*, 19:2003, 2000.
8. D. Fass. met*: a method for discriminating metonymy and metaphor by computer. *Comput. Linguist.*, 17:49–90, March 1991.
9. M. Gedigian, J. Bryant, S. Narayanan, and B. Ciric. Catching metaphors. In *Proceedings of the Third Workshop on Scalable Natural Language Understanding*, ScaNaLU '06, pages 41–48, Stroudsburg, PA, USA, 2006. Association for Computational Linguistics.
10. D. I. Holmes and R. S. Forsyth. The federalist revisited: New directions in authorship attribution. *Literary and Linguistic Computing*, 10(2:111–27, 1995.
11. O. P. John. The "big five" factor taxonomy: Dimensions of personality in the natural language and in questionnaires. *Handbook of personality: Theory and research*, pages pp. 66–100, 1990.
12. V. Keselj, F. Peng, N. Cercone, and C. Thomas. N-gram-based author profiles for authorship attribution, 2003.
13. M. Koppel, S. Argamon, A. R.Shimoni. Automatically categorizing written texts by authors gender. *Literary and Linguistic Computing*, 17(4),pp. 401-412, 2002.
14. H. Liu, R. Mihalcea. Of Men, Women, and Computers: Data-Driven Gender Modeling for Improved User Interfaces, In *Proceedings of the International Conference on Weblogs and Social Media (ICWSM)*, Boulder, Colorado, 2007.
15. R. J. Murphy, S. A. Gray, S. R. Straja, and M. C. Bogert. Student learning preferences and teaching implications. *J Dent Educ.*, 68(8):859–866, 2004.
16. P. M. Niedenthal, S. Krauth-Gruber, F. Ric. Psychology of emotion: interpersonal, experiential, and cognitive approaches. In *Principles of social psychology*, 2006. Psychology Press.

17. B. Pang and L. Lee. A sentimental education: sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, ACL '04, Stroudsburg, PA, USA, 2004. Association for Computational Linguistics.
18. P. Panicheva, J. Cardiff, and P. Rosso. Personal sense and idiolect: Combining authorship attribution and opinion analysis. In N. C. C. Chair), K. Choukri, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis, M. Rosner, and D. Tapias, editors, *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*, Valletta, Malta, may 2010. European Language Resources Association (ELRA).
19. V. Rentoumi, G. Giannakopoulos, V. Karkaletsis, and G. A. Vouros. Sentiment analysis of figurative language using a word sense disambiguation approach. In *Proceedings of the International Conference RANLP-2009*, pages 370–375, Borovets, Bulgaria, September 2009. Association for Computational Linguistics.
20. S. Goswami, S. Sarkar, M. Rustagi. Stylometric Analysis of Blogger's Age and Gender In *Proceedings of AAAI 2009*, 2009.
21. J. Schler, M. Koppel, S. Argamon, J. Pennebaker. Effects of Age and Gender on Blogging. In *Proceedings of AAAI 2006*, 2006.
22. E. Shutova, L. Sun, and A. Korhonen. Metaphor identification using verb and noun clustering. In *Proceedings of the 23rd International Conference on Computational Linguistics*, COLING '10, pages 1002–1010, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics.
23. R. F. Snyder. The relationship between learning styles/multiple intelligences and academic achievement of high school students. *The High School Journal*, 83(2):pp. 11–20, 1999.
24. E. Stamatatos. Ensemble-based author identification using character n-grams. In *3rd International Workshop on Text-based Information Retrieval*, 2006.
25. C. Strapparava and R. Mihalcea. Semeval-2007 task 14: affective text. In *Proceedings of the 4th International Workshop on Semantic Evaluations*, SemEval '07, pages 70–74, Stroudsburg, PA, USA, 2007. Association for Computational Linguistics.
26. P. D. Turney. Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, pages 417–424, Stroudsburg, PA, USA, 2002. Association for Computational Linguistics.
27. T. Wilson, J. Wiebe, and P. Hoffmann. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, HLT '05, pages 347–354, Stroudsburg, PA, USA, 2005. Association for Computational Linguistics.