One System to Solve Them All

Jan Rygl

NLP Centre, Faculty of Informatics, Masaryk University
Czech Republic
rygl@fi.muni.cz

Abstract. People are daily confronted with hundreds of situations in which they could use the knowledge of stylometry. In this paper, I propose a universal system to solve these situations using stylometry features, machine learning techniques and nature language processing tools. The proposed tool can help translation companies to recognize machine translation falsely submitted as a work of a human expert; identify school essays not written by the underwritten student; or cluster product reviews by authors and merge user reviews written by one author using multiple accounts.
All examples above use same techniques and procedures to solve the problem, therefore it is preferred to merge algorithms and implementation of these tasks to a single framework.

Keywords: stylometry, machine learning

1 Introduction

People are daily confronted with hundreds of situations in which they could use the knowledge of stylometry. I will mention several pressing problems:

Purchase of school essays during educational process: With the expansion of the Internet in the majority of households, the number of specialized web pages offering to order essays and diploma theses increased rapidly. If the submitted work was published on the Internet, the plagiarism methods can detect a fraud. Otherwise, stylometry techniques are needed to expose falsely signed works: The style of previous author's works is compared to the style of the submitted work. If the style is different enough that it exceeds the limit defined for the diversity of one author, the system will notify evaluators.

Registering using a false age or gender in dating advertisements; on discussion forums; or in Internet chats: Deception detection is the task of automatically classifying a text as being either truthful or deceptive according to the identity of author such as gender or age. In online social network communities it is easy to provide a false name, age, gender and location in order to hide a true identity, providing criminals such as pedophiles with new possibilities to harm people. Checking user profiles on the basis of text analysis can detect false profiles and flag them for monitoring [6].

Aleš Horák, Pavel Rychlý (Eds.): Proceedings of Recent Advances in Slavonic Natural Language Processing, RASLAN 2014, pp. 19–26, 2014.

Machine translation submitted as human expert translation: Translation companies hire human experts (translators) to translate texts. For some of the less frequented languages it is difficult to verify the quality of the translation, therefore human experts can be tempted to use machine translation tools to complete their tasks automatically. Stylometry techniques can distinguish between automatically translated text and the text translated by a human expert.

False product reviews: During the last five years, the volume of Internet advertising doubled in Czech Republic [11]. Internet shoppers are influenced by product reviews. The share of user reviews increases at the expense of the share of professional reviews. The number of products rises faster than is the capacity of magazines aimed at user reviews. This situation leads to the fact that some companies are guilty of unfair trade practices and creates fake product reviews: positive ones to improve the rating of their goods, and negative ones to harm their competitors [9]. We can fight false reviews by recovering true authorship of reviews; cluster user accounts by their true author; and detect automatically generated reviews.

2 Stylometry

Author's style is defined as a set of measurable text features according to stylostatisticians [8]. These features are called style markers. Word-length frequencies were used as the first style markers to detect an authorship of documents. T. C. Mendenhall discovered that word-length frequency distribution tends to be consistent for one author and differs for different authors (1887, [5]).

Style markers can be divided into categories, which can be defined by properties of texts that are used, or by tools needed to extract information.

Usually, following tool categories are used to implement stylometry techniques (examples of Czech tools are given):

1. Text cleaning (boiler-plate removal, HTML removal, etc.)
2. Language detection
3. Encoding detection (Chared[1])
4. Text tokenization
5. Morphology analysis (Majka [10])
6. Syntactic analysis (SET [4])
7. Semantic analysis (entity detection, abbrevation expansion, etc.)

The number of categories based on extracted information is still growing, therefore only a few predominant examples are listed:

1. Wordclass n-grams
2. Morphology tags n-grams
3. Word-length and sentence-length distribution
4. Typography errors

[1] `http://nlp.fi.muni.cz/projects/chared/`

5. Punctuation usage
6. Subtrees from a tree generated by syntactic analysis

The quality and the utility of style markers depend on the type of problem. Different document lengths and tasks require different style markers, therefore it is recommended to experimentally select a subset of style markers and not to use them all [7].

3 Machine learning

Machine learning techniques work with data instances. Each instance is an n-tuple of features, each feature represents one style marker.

The instances are separated into two groups. Training group is labeled and contains information about author, gender, age, etc., depending on the scenario. Training instances are used to create a machine learning classifier. The classifier is given unlabeled test instances and predicts labels. The features are usually rational numbers, which are automatically normalized to a range $\langle 0, 1\rangle$ or $\langle -1, 1\rangle$.

To solve the problems using stylometry techniques, two Support Vector Machines methods are recommended [3]:

– SVM implementation LIBSVM [1]
– Linear SVM implementation LIBLINEAR [2]

The selected SVM based techniques have several parameters which should be tuned for each data set. Grid search and other optimization techniques are used to find the best parameters for learning data set.

Depending on what is used as a data instance, we can distinguish two approaches 3.1 and 3.2.

3.1 One model per label approach

For each label (labels can be ages, genders, author names, types of translations), one machine learning model is trained. Each model for a label L classifies whether a given document should be given the label L. The n most probable labels are selected for each document ($n = 1$ for a majority of tasks).

The advantage of this approach is that it supports tasks with multilabeled instances (e.g. document written by more authors). The disadvantage is that it requires training instances for each label, therefore this approach cannot predict labels for test instances with unseen labels.

The data instance is an n-tuple of style markers of one document.

3.2 Similarity approach

Similarity approach is used to compare two documents and predict the similarity between them. Given two documents A and B, style-marker n-tuples $s(A)$

and $s(B)$ are extracted. The inverse absolute difference of style-marker n-tuples (similarity) is counted:

$$1 - |s(A)[1] - s(B)[1]|, 1 - |s(A)[2] - s(B)[2]|, \ldots, 1 - |s(A)[n] - s(B)[n]$$

where $s(A)[i]$ is i-th item of $s(A)$ and $s(B)[j]$ is j-th item of $s(B)$.

The data instance is a similarity n-tuple of two style markers. This approach can compare whether two documents have the same label even if the label is not present in training instances.

4 One system

Most of the previously mentioned techniques are common for algorithms solving stylometric problems. Therefore, I proposed a system schema which can be used to solve all tasks with minimal effort. The schema consists of following parts (training a model):

1. Annotating (each doment is given a label)
2. Document processors (documents are cleaned and expanded to a collection of extracted information)
   (a) text cleaning (remove boiler-plae, HTML, ...)
   (b) language detection
   (c) charset detection
   (d) tokenization
   (e) morphology analysis
   (f) syntactic analysis
   (g) semantic analysis (abbreviations, entities, ...)
3. Style extraction (expanded documents are converted to feature n-tuples, where n is the number of style markers)
4. Similarity extraction (if we want to solve a task using a similarity approach, feature n tuples of selected document pairs are compared and similarity n-tuples are counted)
5. Machine learning – training a model (each n-tuple has a label)
   (a) feature selection (select the best combination of style markers)
   (b) machine learning parameters selection
   (c) model creation

The schema for classification consists of the following parts (see Figure 1):

1. Document processors (see a previous List)
2. Style extraction (expanded documents are converted to feature n-tuples, where n is the number of style markers)
3. Similarity extraction (if we want to solve a task using thea similarity approach, feature n tuples of selected document pairs are compared and similarity n-tuples are counted
4. Machine learning classification (each n-tuple is given a label)

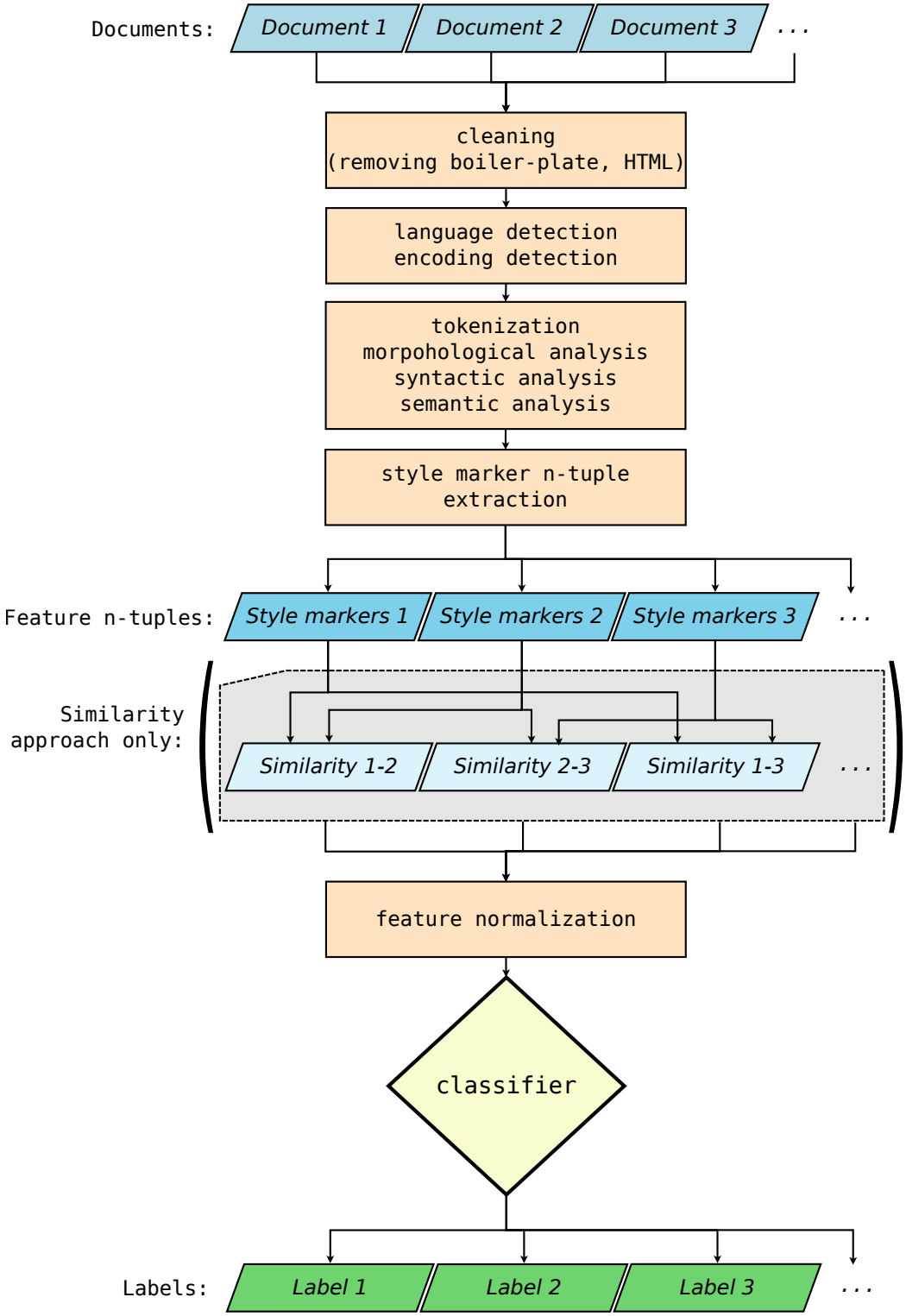

Fig. 1: System schema: document classification

4.1 Authorship of school essays

Input: 2-tuples `(author, document)`.

Style extraction: Select style markers based on genres of documents.

Machine learning: Train a model with two labels (`[author's name]`, `other`) for each author. Author's documents are used as instances with the label `[author's name]`, documents of other authors have the label `other`.

Classification: For each author's model, estimate a probability of each label. Check whether document signed by author `A` has the label `A` with the probability higher than probabilities of all other labels except the `other` label. If author's probability is lower than some probability of other author, notify evaluators.

4.2 False product reviews

Input: 2-tuples `(author, document)`.

Style extraction: Select style markers suitable for short texts.

Similarity extraction: Compare each two documents and extract similarities between them.

Machine learning: Train one model. Comparison of two documents of one author are given a label `same_authors`, pairs of documents signed by different authors are used as instances with a label `different_authors`.

Classification: Check whether pairs consisting of documents from two different authors are labeled as `same_authors`. If more than one document pair of two authors is classified with the same authorship, consider merging these authors.

4.3 Registering using a false age in dating advertisements

Input: 2-tuples `(age, document)`.

Style extraction: Use all style markers.

Machine learning: Divide ages into several groups, each group is represented by one label (e.g. `gradeschooler`, `teen`, `young_adult`). Train one model using these labels.

Classification: Check whether the document is classified as the same label as the document is annotated. If the label does not match, notify system administrators.

4.4 Machine translation submitted as human expert translation

Input: 2-tuples `(source, document)`.

Style extraction: Select style markers based on genres of documents.

Machine learning: Train one model. Machine translation instances are labeled as `machine_translation`, documents translated by human experts have the `human_translation` label.

Classification: Check if the submitted translation is given the `human_translation` label. If the label does not match, evaluate the translation by another human expert.

5 Conclusions and future work

I plan to implement a multilingual system according to the proposed schema. The system will use the state of the art libraries for machine learning techniques and text processing, and wide range of stylometric features. Once implemented, all scenarios mentioned in this paper will be tested using this system and the results will be published.

Acknowledgements This work has been partly supported by the Ministry of Education of CR within the LINDAT-Clarin project LM2010013.

References

1. Chih-Chung Chang and Chih-Jen Lin. *LIBSVM: a library for support vector machines*, 2001. URL: `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.
2. Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin. LIBLINEAR: A Library for Large Linear Classification. *J. Mach. Learn. Res.*, 9:1871–1874, June 2008.
3. Moshe Koppel, Jonathan Schler, and Shlomo Argamon. Computational methods in authorship attribution. *J. Am. Soc. Inf. Sci. Technol.*, 60:9–26, January 2009.
4. Vojtěch Kovář, Aleš Horák, and Miloš Jakubíček. Syntactic analysis using finite patterns: A new parsing system for Czech. In *Human Language Technology. Challenges for Computer Science and Linguistics*, Lecture Notes in Computer Science, pages 161–171, Berlin, 2011. Springer.
5. T. C. Mendenhall. The characteristic curves of composition. *The Popular Science*, 11:237–246, 1887.
6. Claudia Peersman, Walter Daelemans, and Leona Van Vaerenbergh. Predicting age and gender in online social networks.
7. Jan Rygl. Automatic Adaptation of Author's Stylometric Features to Document Types. In *Text, Speech and Dialogue - 17th International Conference. 8655.*, pages 53–61. Brno: Springer, 2014., 2014.

8. Efstathios Stamatatos, Nikos Fakotakis, and George Kokkinakis. Automatic text categorization in terms of genre and author. *Computational Linguistics*, 26(4):471–495, Dec 2000.
9. Ben Verhoeven and Walter Daelemans. Clips stylometry investigation (csi) corpus: A dutch corpus for the detection of age, gender, personality, sentiment and deception in text. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 3081–3085, Reykjavik, Iceland, May 2014. European Language Resources Association (ELRA). ACL Anthology Identifier: L14-1001.
10. Pavel Šmerk. Fast morphological analysis of Czech. In *Proceedings of Third Workshop on Recent Advances in Slavonic Natural Language Processing*, pages 13–16, Brno, 2009. Tribun EU.
11. Association for Internet Advertising. ppm factum, Admosphere, Kantar Media, February 2014. URL `http://www.spir.cz/en/internet-advertising-exceeded-czk-13-billion-last-year-doubling-over-last-five-years`

.