N.N. (Eds.): Proceedings of IJCNLP '11
5th International Joint Conference on Natural Language Processing, Chiang Mai, Thailand, November 8-13

Back to the Roots of Genres: Text Classification by Language Function

Henning Wachsmuth
University of Paderborn, s-lab
Paderborn, Germany
hwachsmuth@s-lab.upb.de

Kathrin Bujna
University of Paderborn
Paderborn, Germany
kabu@mail.upb.de

Abstract

The term "genre" covers different aspects of both texts and documents, and it has led to many classification schemes. This makes different approaches to genre identification incomparable and the task itself unclear. We introduce the linguistically motivated text classification task *language function analysis*, LFA, which focuses on one well-defined aspect of genres. The aim of LFA is to determine whether a text is predominantly expressive, appellative, or informative. LFA can be used in search and mining applications to efficiently filter documents of interest. Our approach to LFA relies on fast machine learning classifiers with features from different research areas. We evaluate this approach on a new corpus with 4,806 product texts from two domains. Within one domain, we correctly classify up to 82% of the texts, but differences in feature distribution limit accuracy on out-of-domain data.

1 Introduction

Text classification has been successfully applied to various natural language processing tasks, among which some of the most popular are topic detection, authorship attribution, sentiment analysis, and genre identification. While the first three refer to single aspects of a text, genres cover different properties of both documents and texts, such as their form, function, purpose, and target audience. As a consequence, many different genre classification schemes exist, which makes most approaches to genre identification badly comparable as a recent study showed (Sharoff *et al.*, 2010). Correspondingly, the question which features work best in genre identification still remains open. We argue that one major reason behind is a missing common understanding of genres and that we need to focus on the single aspects of genres in order to overcome this situation.

In this paper, we investigate *why* a text was written. Therefore, we introduce the ambitious task *language function analysis*, abbreviated as LFA, i.e., to classify the predominant function of a text as intended by its author. In line with the work of the psychologist Karl Bühler (1934), we distinguish three abstract and very general classes, namely, *expressive*, *appellative*, and *informative* texts. Several search and text mining applications can benefit from applying LFA as a document filtering step with respect to both efficiency and effectiveness. A search engine, for instance, might restrict its result list to hits that mainly serve an informative purpose. Similarly, LFA can be used in opinion mining to cancel out promotional texts in favor of personal attitudes. While, of course, LFA does not replace genre identification, we think that language functions constitute one root of a common and clear genre concept.

Language functions are well-studied in linguistic pragmatics, but we analyze whether they also correlate with statistical text characteristics. For this purpose, we built a manually annotated corpus with 4,806 product-related texts from two separated domains (music and smartphones) in close collaboration with industry. Each text is tagged as *personal*, *commercial*, or *informational*, which can be seen as an application-specific classification by language function. Also, the texts have been categorized by sentiment polarity.

Our approach to LFA relies on machine learning of lexical and shallow linguistic features from different research areas. We evaluate this approach both for classification within one corpus domain and for the transfer to another domain. With respect to the in-domain task, our results indicate that a text collection of homogeneous quality and style allows for high accuracy. In particular, we

correctly classify 81.9% of the music texts using a very efficiently computable feature set. This makes our approach suitable for document filtering purposes. However, classification of out-of-domain data seems difficult because of the *covariate shift* (Shimodaira, 2000) in feature distribution between domains. Interestingly, though, the best-performing features for this task come from the area of authorship attribution.

1.1 Summary of Contributions

Altogether, the main contributions of this paper are the following:

- We introduce the linguistically motivated text classification task language function analysis, which addresses one well-defined aspect of genres (Section 2).
- We provide a corpus for language function analysis and sentiment analysis with product-related texts that were manually annotated by one of our industrial partners (Section 4).
- We analyze the impact of machine learning features from different research areas on the language function analysis of texts from two domains (Sections 5 and 6).

2 Language Function Analysis

One of the most influential attempts to categorize language functions was introduced by the famous psychologist Karl Bühler (1934). In his *Organon model*, which is rooted in Plato's view of language as a tool, Bühler identifies and interrelates three fundamental functions of natural language in communication: the *expression* of the speaker, the *appeal* to the receiver, and the *representation* of the object or state of affair being communicated. As illustrated in Figure 1 they all refer to a linguistic sign, which can be understood as the unit of all forms of language.

Based on the three language functions, Katharina Reiß (1971) defined a classification of text types, which relates to the intention of the author of a text. In particular, she distinguished between the form-focused expression of the author's attitudes in *expressive* texts, the aim of making an appeal to the reader in *appellative* (or *operative*) texts, and the content-focused description of objects and facts in *informative* texts. Reiß assigned several concrete text types such as "report" (informative), "novel" (expressive), or "comment" (informative and appellative) to one or more of these classes. While she claimed that a hybrid type is the regular case, she observed that one function is predominant in most texts. We adopt Reiß' typology to define the language function analysis task.

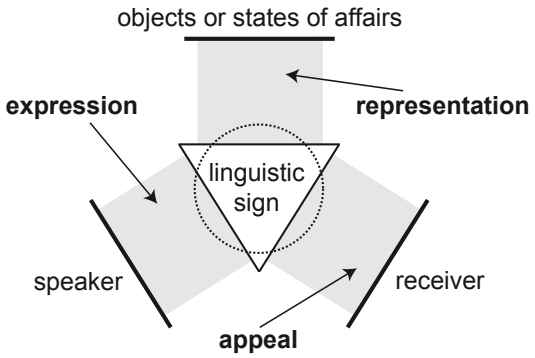

Figure 1: The organon model, formulated by Karl Bühler (1934), with its three language functions expression, appeal, and representation.

Definition 1 (Language Function Analysis) *Let $C = \{$expressive, appellative, informative$\}$ be the set of abstract language functions and let d be a text. Then the task of language function analysis, LFA, is to find the mapping $d \mapsto c \in C$ such that c is the predominant language function of d.*

We argue that LFA can help in many practical problems where document filtering is needed, especially because of its generic nature. For product-related texts, the use of LFA emerges, when we map the abstract functions in C to the following concrete language function classes of text:

- *personal (expressive).* Text that aims to express the personal attitude of an individual towards a product of interest.
- *commercial (appellative).* Text that follows commercial purposes with respect to a product of interest.
- *informational (informative).* Text that reports on a product of interest in an objective and journalistic manner.

Another example for a set of concrete language function classes might be *review* (expressive), *proposal* (appellative), and *report* (informative) in the research project context. Notice, though, that the mapping from abstract functions to concrete classes of text is meant to be an interpretation for a concrete learning situation rather than a redefinition of the task. In the remaining sections, we

use the classes *personal*, *commercial*, and *informational* for a first evaluation of LFA. Our intuition is that, statistically, language functions imply shallow linguistic features, such as certain parts-of-speech or writing style characteristics.

3 Related Work

Classification by intention has been recently addressed in (Kröll and Strohmaier, 2009). The authors infer a subset of 145 intentions from the transcriptions of speeches based on actions mentioned *within* the text. Similar problems refer to the analysis of speaker intentions in conversation (Kadoya *et al.*, 2005) and to the area of textual entailment (Michael, 2009), which is about the information implied by text. In contrast, LFA is about the question why a text was written and, thus, refers to the authorial intention *behind* a text.[1]

In that, our work resembles (Santini, 2005) where a linguistic expert system analyzes the gradation of four text types that convey the purpose and function of a text. Two of these types fit to the abstract functions introduced in Section 2, namely, the types "explicatory/informational" and "argumentative/persuasive". However, the other types ("descriptive/narrative" and "instructional") seem quite arbitrary and appear to intersect with the first type. Moreover, a class for the expression of personal views is missing. This might result from the used SPIRIT corpus (Clarke et. al., 2002), which was created for question answering purposes. Besides, language functions constitute a general classification scheme that may be concretized for a task at hand, while Santini regards text types only as input to web genre identification.

In terms of the communicative purpose of a text, language functions can be considered as the most abstract view on genres. Indeed, Kessler *et al.* (1997) argue that the class of texts that aim at persuading someone would not be seen as a genre merely because that class is too generic. While the authors simply define a genre to refer to a certain functional trait as well as to some formal properties, several abstract and concrete classification schemes have been proposed for genre identification. In a pioneer study on genres, Biber (1986) analyzes basic textual dimensions, such as "informative vs. involved", while Karlgren and Cutting (1994) try to automatically separate informative from imaginative texts. Stamatatos *et. al.* (2000) rely on more concrete press-related genres (e.g. "Letter to the editor" or "Spot news"), and Garera and Yarowsky (2009) investigate modern conversational genres, such as "Email".

The two latter show that genres do not only describe the function and purpose of a text, but also its form and target audience and, thus, also represent concepts orthogonal to language functions. Correspondingly, a great deal of genre research in the last decade focused on web genres as surveyed in (Stein *et al.*, 2010). Two standard corpora for web genre identification, KI-04 (Meyer zu Eissen and Stein, 2004) and SANTINIS (Santini, 2010), illustrate a common situation in genre research: Their classification schemes partly overlap, e.g. the class "Help" from KI-04 can be mapped to "FAQ" in SANTINIS, but partly also contain unique and quite specific classes, such as "Search pages" in SANTINIS. Moreover, Sharoff *et. al.* (2010) found out that seemingly similar classes ("Portrayal" and "Personal home page") differ strongly in terms of discriminative features. This supports our argumentation that there is neither a clear common understanding of genres, nor a well-defined genre concept. Additionally, Boese and Howe (2005) recognized that, in the web, genres may evolve over time to other genres.

However, language functions still represent one important aspect of genres. Accordingly, genre identification and LFA have similarities with respect to both practical applications (e.g. document filtering) and potentially helpful features. Since our focus is on a text itself as opposed to a document, we follow Webber (2009) who emphasizes the importance of text-internal and linguistic features, such as particular parts-of-speech. Also, we investigate both character-based and word-based n-grams, which were most successful in the above-mentioned evaluation of genre collections of Sharoff *et al.* (2010).

Further promising features relate to sentiment analysis and authorship attribution. Like LFA, sentiment analysis covers the issue of subjectivity (Pang and Lee, 2008), but it addresses *what* is said. Correspondingly, research in sentiment analysis often focuses only on characteristic terms as in (Prettenhofer and Stein, 2010). In contrast, approaches to authorship attribution aim at measuring the writing style of a text; sometimes based on

[1] While literature theory also addresses the intention of the reader and the intention of the text itself (Eco, 1990), only authorial intention is relevant for the purpose of this paper.

lexical and shallow linguistic information (Luyckx and Daelemans, 2008), sometimes using deeper analyses like parsing (Raghavan *et al.*, 2010). We adopt some of these features in Section 5 and 6.

4 The LFA-11 Corpus

To evaluate LFA, we built the LFA-11 corpus with manually annotated German texts from two separated domains: *music* and *smartphones*. The purpose of the corpus is to provide textual data for the development and evaluation of approaches to LFA and sentiment analysis. The corpus is freely available at http://infexba.upb.de.

The music collection of LFA-11 contains 2,713 promotional texts, professional and user reviews that were taken from a social network platform. Accordingly, these texts are well-written and of homogeneous style. In contrast, a set of 2,093 blog posts from the *Spinn3r corpus*[2] addresses smartphones. The Spinn3r project aims at crawling and indexing the whole blogosphere. Hence, the texts in the smartphone collection vary strongly in quality and writing style. While the music texts span 9.4 sentences with 23.0 tokens on average, the blog posts have an average length of 11.8 sentences but only 18.6 tokens per sentence.

4.1 Annotations

The corpus consists of UTF-8 encoded XMI files preformatted for the Apache UIMA framework[3], which implements the *Unstructured Information Management Architecture* (Ferrucci and Lally, 2004). Each file includes the text together with one of the language function annotations *personal*, *commercial*, and *informational*. Also, the texts have been classified by sentiment polarity as positive, negative, or neutral. Tagging was done by two employees of the *Digital Collections Verlagsgesellschaft mbH*, a leading supplier of digital asset management systems.

Figure 2 shows excerpts from three texts of the music collection, one out of each language function class. The excerpts have been translated to English for clarity. While some indicators of language functions might have been lost due to translation, the examples underline the strong connection of the concrete language function classes to the abstract functions from Section 2. In order to support a consistent categorization, the following guidelines were given to the two employees for the language function annotations:

personal. *... How did Alex recently ask when he saw Kravitz' latest best-of collection: Is it his own liking, the voting on his website, or the chart position what counts? Good question. However, in our case, there is nothing to argue about: 27 songs, all were number one. The Beatles. Biggest band on the globe. ...*

commercial. *... The sitars sound authentically Indian. In combination with the three-part harmonious singing and the jingle-jangle of the Rickenbacker guitars, they create an oriental flair without losing their Beatlesque elegance. If that doesn't make you smile! ...*

informational. *... "It's All Too Much"? No, no, still okay, though an enormous hype was made for decades about the seemingly new Beatles song. The point is that exactly this song "Hey Bulldog" has already been published long time ago, most recently on a reprint of "Yellow Submarine" in the year 1987. ...*

Figure 2: Translated excerpts from three texts of the music collection. Note that the translation to English might have affected the indicators of the corresponding language functions.

- *personal.* "Use this annotation if the text seems not to be of commercial interest, but probably represents the personal view on the product of a private individual."
- *commercial.* "Use this annotation if the text is of obvious commercial interest. The text seems to predominantly aim at persuading the reader to buy or like the product."
- *informational.* "Use this annotation if the text seems not to be of commercial interest with respect to the product. Instead, it predominantly appears to be informative in a journalistic manner."

About 20% of the music texts and 40% of the smartphone texts were tagged twice in order to compute inter-annotator agreement. The resulting values κ_m = 0.78 (music) and κ_s = 0.67 (smartphone) of Cohen's Kappa (Carletta, 1996) for the language function annotations constitute "substantial agreement". Especially κ_s is far from perfect, which can be problematic for text classification purposes. Under consideration of the hybridity of language functions in texts (cf. Section 2), κ_m and κ_s appear to be quite high, though.

[2] Spinn3r corpus, http://www.spinn3r.com
[3] Apache UIMA, http://uima.apache.org

Set	personal	commercial	informational
music collection			
Training	521 (38.5%)	127 (9.4%)	707 (52.2%)
Validation	419 (61.7%)	72 (10.6%)	188 (27.7%)
Test	342 (50.4%)	68 (10.0%)	269 (39.6%)
smartphone collection			
Training	546 (52.1%)	90 (8.6%)	411 (39.3%)
Validation	279 (53.3%)	36 (6.9%)	208 (39.8%)
Test	302 (57.7%)	28 (5.4%)	193 (36.9%)

Table 1: Distribution of language function classes in the music and smartphone sets of the corpus.

Set	positive	neutral	negative
music collection			
Training	1003 (74.0%)	259 (19.1%)	93 (6.9%)
Validation	558 (82.2%)	82 (12.1%)	39 (5.7%)
Test	514 (75.7%)	115 (16.9%)	50 (7.4%)
smartphone collection			
Training	205 (19.6%)	738 (70.5%)	104 (9.9%)
Validation	110 (21.0%)	343 (65.6%)	70 (13.4%)
Test	84 (16.1%)	359 (68.6%)	80 (15.3%)

Table 2: Distribution of sentiment polarity classes in the music and smartphone sets of the corpus.

4.2 Evaluation Sets

We created splits for each domain with half of the texts in the training set and each one fourth in the validation set and test set, respectively. Table 1 and Table 2 show the class distributions of language functions and sentiment polarities. The distributions indicate that the training, validation, and test sets differ significantly from each other. Also, Table 1 and 2 give a first hint that the correlation between language functions and sentiment is low. With regard to the distribution of language functions, we observe a large imbalance between the three classes. In case of double-annotated texts, we chose the annotations of the employee who categorized more texts as *commercial*. Still, this class remains by far the minority class with only about 5% to 10% of the texts in all sets. This, of course, makes the task at hand more difficult.

5 Features

To investigate whether LFA has correlations with other text classification tasks, we experimented with several lexical and shallow linguistic features that relate to some of the research areas mentioned in Section 3. For a concise evaluation, we organized these features into the following six types.

1. *Simple genre features.* Simple approaches from genre identification: the frequency of each part-of-speech tag that occurs at least 15 times in the training set, the average word length, and the *Lix* readability index (Anderson, 1983).

2. *Text type.* Features inspired by linguistic expert knowledge from (Santini, 2005), namely, the frequency of time and money entities as well as the frequency of two sets of part-of-speech tags: a) personal and possessive pronouns, b) nouns and adjectives.

3. *Writing style.* A selection of measures used in authorship attribution (Stamatatos, 2009): the frequency of the most common words, part-of-speech trigrams, and character trigrams as well as the frequency of capitalized, upper-case, and lower-case words. The same for parentheses, punctuation and quotation marks, and the portion of "?" and "!" under all sentence delimiters.

4. *Sentiment.* Indicators for sentiment, namely, the frequency of 15 common emoticons such as ";-)" and the sentiment polarity of the text.

5. *Core trigrams.* The frequency of the most discriminative part-of-speech and character trigrams of each language function class. Similar features performed best in (Sharoff *et al.*, 2010). Here, we use all trigrams that occur over six times as often in one class c as in any other class $c' \neq c$.

6. *Core vocabularies.* The frequency of the most discriminative words, introduced as a genre feature by Lee and Myaeng (2002). We define such a word to occur at least in 15 training texts of class c and over six times as often in c as in any other class $c' \neq c$.

6 Experiments

We now report on an evaluation of our approach to text classification by language function. As text classification often suffers from domain dependency, i.e., effective results are only achieved in the learning domain, we experimented with both corpus domains. The goal of our evaluation can be seen as three-fold: first, to evaluate the effectiveness of an LFA classifier on in-domain and out-of-domain data, second, to analyze the impact of each single feature type from Section 5, and third,

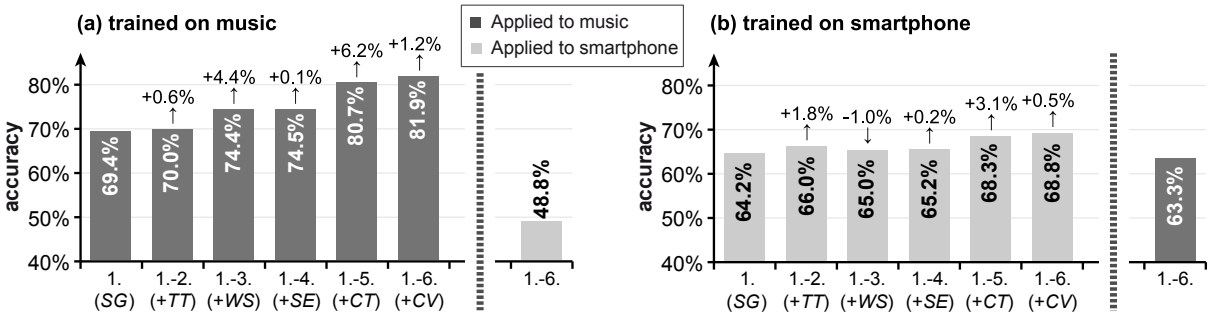

Figure 3: Classification accuracy in the LFA task for a stepwise integration of the six feature types and the transfer to another domain: (a) Training on the music training set, application to the music test set, and transfer to the smartphone test set. (b) Training on the smartphone training set, application to the smartphone test set, and transfer to the music test set.

to check whether LFA based on supervised learning qualifies for document filtering purposes.[4]

6.1 Experimental Set-up

Since commercial texts are largely underrepresented in the music and the smartphone collection, the two training sets were balanced with oversampling. After sentence splitting and tokenization, we applied the highly efficient *TreeTagger* (Schmid, 1995) for part-of-speech tagging and we extracted time and money entities with fast regular expressions. Regarding sentiment polarity, we used the corpus annotations for simplicity.

For the writing style features, we determined the 48 most common words, the 55 most common part-of-speech trigrams, and the 35 most common character trigrams on the training set of each collection. Accordingly, we computed the most discriminative trigrams for feature type 5. The core vocabularies sum up to 30 words for the music collection and to 36 words for the smartphone collection (proper names were discarded). Some of these words are quite specific, e.g. "single" in the music domain (an indicator for *commercial*), while others seem less domain-dependent such as "zumindest" ("at least", *informational*).

Altogether, the six feature types were instantiated by 299 music and 373 smartphone features, respectively. On both training sets, we trained one linear multi-class support vector machine (hereafter called SVM) using feature type 1 to m for each $m \in [1, 6]$ as well as one such SVM using only type m. For this, we applied the *LibSVM* integration in *Weka* (Hall et. al., 2009; Fan et. al., 2001), where we selected the cost parameters of all SVMs on the validation sets. Finally, we analyzed the impact of the resulting classifiers on both test sets. We measured their effectiveness in terms of accuracy, precision, and recall and their efficiency in milliseconds per processed input text.

[4]The Java source code and the feature files used for evaluation can be accessed at http://infexba.upb.de.

6.2 Results

Effectiveness of the classifiers. Figure 3a illustrates classification accuracy on the music test set for a stepwise integration of feature type 1 to 6 into an SVM trained on the music training set. Additionally, the accuracy for the transfer of the SVM with all features to the smartphone domain is depicted. The simple genre features (*SG*) already achieved 69.4%. While text type (*TT*) and sentiment (*SE*) contributed only little, the writing style features (*WS*) and the core trigrams (*CT*) boosted accuracy by 4.4% and 6.2%, respectively. At last, the core vocabularies (*CV*) added 1.2 percentage points to the resulting overall accuracy of 81.9% (86.7% on the validation set).

When we applied the SVM with all features to the smartphone test set, its accuracy dropped to 48.8%. To find out whether this dramatic decrease indicates a covariate shift in the feature distribution between the two domains, we retrained the SVM on the smartphone training set. Indeed, its accuracy was re-increased to 68.8%, as shown in Figure 3b. Interestingly, though, the domain transfer worked fine in the opposite direction, i.e., the SVM trained on smartphone texts still correctly classified 63.3% of the texts in the music test set. We suppose that this effect originates from the heterogeneity of the smartphone texts, which prevented the learning of features from being biased towards a certain style of speech. Such a bias naturally exists in music reviews and the like.

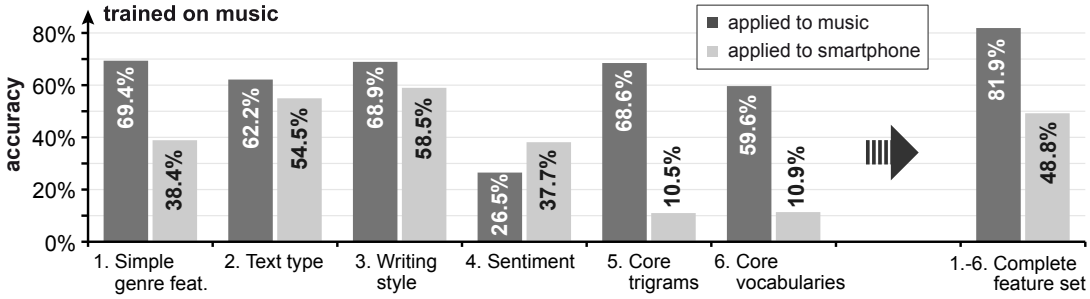

Figure 4: Classification accuracy of each feature type trained on the music training set and applied to both the music test set and the smartphone test set. The accuracy of all features is given on the right.

Domain	Class	Precision	Recall	F_1
music	personal	88.7%	84.8%	86.7%
	commercial	61.9%	88.2%	72.7%
	informational	80.8%	76.6%	78.6%
smartphone	personal	83.5%	68.5%	75.3%
	commercial	23.2%	46.4%	31.0%
	informational	63.9%	72.5%	68.0%

Table 3: Precision, recall and F_1-score for the three language function classes on the in-domain test sets using the SVM with all six feature types.

Nevertheless, Figure 3b also conveys that we achieved 13.1% less accuracy in the smartphone domain than in the music domain (68.8% vs. 81.9%). The accuracy of *SG*, 64.2%, was over five points lower, and only the integration of *TT* (+1.8%) and *CT* (+3.1%) yielded notable improvements afterwards. Adding *WS* even led to a decrease of one percentage point. However, this does not mean that the writing style features failed, as we see later on, but seems to be only noise from the optimization process of the SVM.

While a kappa value of 0.67 (cf. Section 4) renders high accuracy difficult, in general, one reason for the weak performance on the smartphone test set can be inferred from Table 3. This table lists effectiveness results of the SVMs with all features for each class. On the music test set, we observe a recall of more than 75% for all three classes. Though precision significantly dropped for *commercial*, given a class imbalance of 1:9 (commercial:rest), 61.9% is still over five times better than the expected precision of guessing. In contrast, the recognition of commercial texts failed on the smartphone test set with an F_1-score of only 31.0%. Apparently, the SVM did not determine meaningful features for *commercial*, probably because of the small number of commercial texts (cf. Table 1) in combination with the heterogeneity of the smartphone collection. This, in turn, also affected the effectiveness of the other classes.

Effectiveness of the feature types. We measured the effectiveness of each feature type in isolation in order to investigate their impact on LFA. Within music, the simple genre features, the writing style type, and the core trigrams did best, each of them with nearly 70% accuracy as shown in Figure 4. However, there is not *one* discriminative type, i.e., the complete feature set clearly outperformed all single types. Under the transfer to the smartphone domain, only the text type and writing style features reasonably maintained effectiveness. The core vocabularies failed on out-of-domain data, and also the core trigrams did unexpectedly bad, dropping from 68.6% to 10.5%.

With regard to sentiment, our evaluation underpins the observation from Section 4 that sentiment polarities and language functions hardly correlate: the according features learned on the music training set did not work out on both test sets, and the same holds for the opposite direction in Figure 5. There, we see that feature type 1, 3, and 5 also performed best within the smartphone domain. In particular, the writing style type (64.8%) is only 4% worse than the complete feature set. Moreover, while the domain transfer worked well in general, again the most impact was achieved by the text type features with 59.8% accuracy and by the writing style features with 58.9%. This suggests that the distribution of these features is only weakly domain-dependent in LFA. With respect to the writing style type, this is an interesting result, as it indicates that the genre-related task LFA significantly benefits from features that are most prominent in the area of authorship attribution.

Efficiency. We measured the run-time of the classifier with the complete feature set ten times on the music test set using a *2 GHz Intel Core*

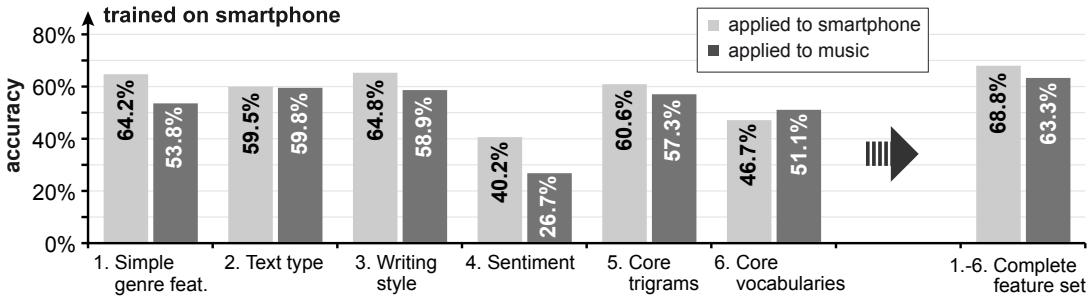

Figure 5: Classification accuracy of each feature type trained on the smartphone training set and applied to both the smartphone test set and the music test set. The accuracy of all features is given on the right.

2 Duo MacBook with *4 GB RAM*. Including all processing steps (sentence splitting, tokenization, part-of-speech tagging, entity recognition, feature extraction, and classification), the average runtime was 37.0 ms per text (σ = 0.6 ms) compared to 6.2 ms needed for tokenization alone. At least for the homogeneous music domain, where we achieved high accuracy, we thus claim that our approach is suitable for fast document filtering.

7 Conclusion

We presented the text classification task language function analysis, LFA, which addresses *why* a text was written and which is motivated by Karl Bühler's functions of natural language. We see language functions as one root of a well-defined genre concept and we argue that a common understanding of such a concept is needed in order to achieve real progress in genre research.

For evaluation of LFA, we provide the LFA-11 corpus with product-related texts from two very different domains that was developed in collaboration with industry. Each text in the corpus has been manually classified by its concrete language function. Approaching LFA with machine learning, we achieved promising results within one homogeneous domain. Moreover, we found out that features commonly used in authorship attribution have the most impact on LFA in both evaluated domains and that they also qualify for domain transfer. This indicates that language functions relate to the writing style of a text. In contrast, the correlation with sentiment appeared to be low.

However, in general, both the language function analysis of more heterogeneous texts and the domain transfer remain unsolved. In future work, we hence aim to investigate the use of sentence-level classification and domain adaptation techniques to further improve our approach to LFA.

Acknowledgments

This work was partly funded by the German Federal Ministry of Education and Research (BMBF) under contract number 01IS08007A.

References

Jonathan Anderson. 1983. Lix and Rix: Variations on a Little-known Readability Index. *Journal of Reading*, 26(6):490–496.

Douglas Biber. 1986. Spoken and Written Textual Dimensions in English: Resolving the Contradictory Findings. *Language*, 62(2):384–413.

Elizabeth S. Boese and Adele E. Howe. 2005. Effects of Web Document Evolution on Genre Classication. In *Proceedings of the 14th ACM International Conference on Information and Knowledge Management*, pages 639–646, Bremen, Germany.

Karl Bühler. 1934. *Sprachtheorie. Die Darstellungsfunktion der Sprache*. Verlag von Gustav Fischer, Jena, Germany.

Jean Carletta. 1996. Assessing Agreement on Classification Tasks: The Kappa Statistic. *Computational Linguistics*, 22: 249–254.

Charles L. A. Clarke, Gordon V. Cormack, M. Laszlo, Thomas R. Lynam, and Egidio L. Terra. 2002. The Impact of Corpus Size on Question Answering Performance. In *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 369–370, Tampere, Finland.

Umberto Eco. 1990. *I Limiti dell'Interpretazione*. Bompiani, Milano, Italy.

Rong-En Fan, Pai-Hsuen Chen, and Chih-Jen Lin. 2001. Working Set Selection Using Second Order Information for Training Support Vector Machines. *Journal of Machine Learning Research*, 6:1889–1918.

David Ferrucci and Adam Lally. 2004. UIMA: An Architectural Approach to Unstructured Information

Processing in the Corporate Research Environment. *Natural Language Engineering*, 10(3–4):327–348.

Nikesh Garera and David Yarowsky. 2009. Modeling Latent Biographic Attributes in Conversational Genres. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the Association for Computational Linguistics and the 4th International Joint Conference on Natural Language Processing*, pages 710–718, Suntec, Singapore.

Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. 2009. The WEKA Data Mining Software: An Update. *SIGKDD Explorations*, 11(1).

Yuki Kadoya, Kazuhiro Morita, Masao Fuketa, Masaki Oono, El-Sayed Atlam, Toru Sumitomo, and Jun-Ichi Aoe. 2005. A Sentence Classification Technique using Intention Association Expressions. *International Journal of Computer Mathematics*, 82(7):777–792.

Jussi Karlgren and Douglass Cutting. 1994. Recognizing Text Genres with Simple Metrics using Discriminant Analysis. In *Proceedings of the 15th International Conference on Computational Linguistics*, pages 1071–1075, Kyoto, Japan.

Brett Kessler, Geoffrey Numberg and Hinrich Schütze. 1997. Automatic Detection of Text Genre. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics*, pages 32–38, Madrid, Spain.

Mark Kröll and Markus Strohmaier. 2009. Analyzing Human Intentions in Natural Language Text. In *Proceedings of the Fifth International Conference on Knowledge Capture*, pages 197–198, Redondo Beach, CA.

Yong-Bae Lee and Sung Hyon Myaeng. 2002. Text Genre Classification with Genre-Revealing and Subject-Revealing Deatures. In *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 145–150, Tampere, Finland.

Loizos Michael. 2009. Reading between the Lines. In *Proceedings of the 21st International Joint Conference on Artifical intelligence*, pages 1525–1530, San Francisco, CA.

Kim Luyckx and Walter Daelemans. 2008. Authorship Attribution and Verification with Many Authors and Limited Data. In *Proceedings of the 22nd International Conference on Computational Linguistics*, pages 513–520, Manchester, UK.

Sven Meyer zu Eissen and Benno Stein. 2004. Genre Classification of Web Pages: User Study and Feasibility Analysis. In *Proceedings of the 27th German Conference on Artificial Intelligence*, pages 256–269, Ulm, Germany.

Bo Pang and Lillian Lee. 2008. Opinion Mining and Sentiment Analysis. *Foundations and Trends in Information Retrieval*, 2(1–2):1–135.

Peter Prettenhofer and Benno Stein. 2010. Cross-Language Text Classification using Structural Correspondence Learning. In *Proceedings of the 48th Annual Meeting of the Association of Computational Linguistics*, pages 1118-1127, Uppsala, Sweden.

Sindhu Raghavan, Adriana Kovashka, and Raymond Mooney. 2010. Authorship Attribution using Probabilistic Context-Free Grammars. In *Proceedings of the ACL 2010 Conference Short Papers*, pages 38–42, Uppsala, Sweden.

Katharina Reiß. 1971. *Möglichkeiten und Grenzen der Übersetzungskritik*. Max Hueber Verlag, Munich, Germany.

Marina Santini. 2005. Automatic Text Analysis: Gradations of Text Types in Web Pages. In *Proceedings of the Tenth ESSLLI Student Session*, pages 276–285, Edinburgh, UK.

Marina Santini. 2010. Cross-testing a Genre Classication Model for the Web. *Genres on the Web: Computational Models and Empirical Studies*, Springer.

Serge Sharoff, Zhili Wu, and Katja Markert. 2010. The Web Library of Babel: Evaluating Genre Collections. In *Proceedings of the Seventh Language Resources and Evaluation Conference, LREC 2010*, pages 3063–3070, Malta.

Hidetoshi Shimodaira. 2000. Improving Predictive Inference under Covariate Shift by Weighting the Log-Likelihood Function. *Journal of Statistical Planning and Inference*, 90(2):227–244.

Helmut Schmid. 1995. Improvements in Part-of-Speech Tagging with an Application to German. In *Proceedings of the EACL SIGDAT-Workshop*, pages 47–50, Dublin, Ireland.

Efstathios Stamatatos, Nikos Fakotakis, and George K. Kokkinakis. 2000. Text Genre Detection using Common Word Frequencies. In *Proceedings of the 18th International Conference on Computational Linguistics*, pages 808–814, Saarbrücken, Germany.

Efstathios Stamatatos. 2009. A Survey of Modern Authorship Attribution Methods. *Journal of the American Society for Information Science and Technology*, 60(3):538–556.

Benno Stein, Sven Meyer zu Eißen, and Nedim Lipka. 2010. Web Genre Analysis: Use Cases, Retrieval Models, and Implementation Issues. *Text, Speech and Language Technology*, 42:167–190.

Bonnie Webber. 2009. Genre Distinctions for Discourse in the Penn TreeBank. In *Proceedings of the 47th Annual Meeting of the Association for Computational Linguistics and the 4th International Joint Conference on Natural Language Processing*, pages 674–682, Suntec, Singapore.