

Feature Instability as a Criterion for Selecting Potential Style Markers

Moshe Koppel
Navot Akiva
Ido Dagan

{koppel, navot, dagan}@cs.biu.ac.il
Computer Science Department, Bar Ilan University, Ramat Gan 52900, Israel

Abstract

We introduce a new measure on linguistic features, called *stability*, which captures the extent to which a language element, such as a word or a syntactic construct, is replaceable by semantically equivalent elements. This measure may be perceived as quantifying the degree of available “synonymy” for a language item. We show that frequent but unstable features are especially useful as discriminators of an author’s writing style.

1. Introduction

Often we wish to find linguistic markers that distinguish the writing style of a particular author or class of authors. We seek features that are typically used variably by different authors but are used consistently by any given author – or, at least, by the author whose writing we wish to distinguish. Obviously, these markers will differ from author to author.

In this paper, we wish to identify the pool of potentially useful linguistic features from which markers might fruitfully be chosen. To be precise, we do not seek features that distinguish a particular author but rather author-independent criteria for ranking features which are worth considering when seeking distinguishing characteristics of *any* given author.

Consider some examples. If a particular author were found to use the word *awful* more frequently than other authors, this would certainly be worth noting. After all, the word *bad* is generally a reasonable, and more common, alternative to *awful*. The fact that our author chooses to use the word *awful* frequently therefore likely reflects deliberate stylistic choice that we might profitably exploit for identifying the author’s writing. What about the word *touchdown*? The frequency of use of *touchdown* is a rather poor feature for identifying an author’s style. There are actually two distinct reasons that this is so. First, there is no alternative word for expressing the concept *touchdown* so that its use does not reflect stylistic choice. Second, *touchdown* is tightly tied to a particular topic so that its frequency of use in one corpus

is unlikely to reflect the frequency with which it will be used in other documents, which might concern other topics. By contrast, *awful* is commonly used across most topics and has plausible alternatives. Some words satisfy one of these two criteria but not both. For example, *perspire* offers a plausible alternative (*sweat*) but is not frequently used across a broad range of topics. On the other hand, words like *blue* and *ten* are used across topics but don't have common alternatives.

To summarize, we seek linguistic features the use of which might reflect deliberate stylistic choice by a given author. Such features will tend to be used across topics and will offer plausible linguistic alternatives. The first criterion is relatively easy to approximate by checking overall frequency of use across different topic areas. We will focus on a formal definition of the second criteria: how can we determine the extent to which a particular linguistic feature offers alternatives? Although the examples we gave are all words, the criteria we define will extend beyond words to other types of linguistic features.

2. Linguistic Feature Stability

Data-driven approaches to text and language processing apply various quantitative measures to linguistic elements, such as words and terms. These measures capture important properties of language items and are often utilized in various ways for different computational tasks, such as information retrieval, text classification, terminology extraction and statistical thesaurus construction.

In this paper, we introduce a new type of measure for linguistic elements that we call *meaning-preserving stability*, or *stability* for short. It captures the degree to which a linguistic element or construct can be substituted for in a piece of text without affecting the text meaning. This measure is applied most intuitively to words and terms, but is applicable also to other types of linguistic constructs, such as part-of-speech sequences. Stability captures interesting properties in language and thus seems interesting from a purely scientific point of view, but it also has potential use within applied tasks. Stability can be perceived as quantifying the typical degree of available “synonymy” for a language element, while generalizing the notion of synonymy to any type of linguistic feature rather than being restricted to its common use for words only.

To make this a bit more concrete, let us begin with a simple example. Consider the three sentences

1. *John was lying on the couch next to the window.*
2. *John was reclining on the sofa by the window.*
3. *John had been lying on the couch near the window.*

The three sentences convey (approximately) the same message. Some of the words remain invariant in all three versions (*John, window*), while others are replaced by other words (*was : had been, couch : sofa, next to : by : near, lying : reclining*) which don't significantly change the meaning of the sentence. More generally, consider features of a text such as word or phrase frequencies, frequencies of syntactic structures or any other feature whose value can be measured in a given text. Roughly speaking, the *stability* of such a feature is the extent to which the measured value of that feature tends to remain invariant across different texts that convey the same meaning. For example, proper nouns are very stable, while words with many synonyms are unstable.

Obtaining training material for measuring stability is a difficult challenge, requiring alternative versions of texts that carry – in the ideal case, exactly – the same meaning. “Parallel” (monolingual) texts of this nature have been used for automatic paraphrase extraction, using either multiple translations of the same text or news stories from different sources that describe the same event [Barzilay and McKeown 2001; Shinyama et al. 2002]. The disadvantage of these types of training material is that they rely on the fact that different people have manually created different versions of the same text. Such manually created versions are not easily available for most texts. In order to lighten the supervision requirements, we have used a machine translation (MT) system to translate our training corpus back-and-forth via multiple languages, thus obtaining several English versions of the same text. Indeed, the use of automatic translation to generate training data is problematic for several reasons, as discussed below. Yet, the combination of MT with other technologies that can leverage the stability measure, such as style-based text categorization, provides appealing advantages that reduce the overall dependency on particular training materials.

In the remainder of the paper, we first present the general notion of stability and define a concrete stability measure. We then present empirical measurements of stability and illustrate the properties that it captures. Finally, we demonstrate how stability can be utilized effectively for feature selection in style-based text classification.

3. A Stability Measure

A stability *measure* would be a quantitative measure that correlates with a feature's tendency to be preserved across different meaning-preserving variants of a text. We formally define a specific stability measure and consider empirical results of stability experiments on the Reuters 21578 corpus [Lewis, 1997] and on the British National Corpus (BNC).

Let $\{d_1, d_2, \dots, d_n\}$ be a set of documents (or text segments) and let $\{d_i^1, d_i^2, \dots, d_i^m\}$ be $m > 1$ different versions of d_i

where the meanings of the m versions are all roughly identical. For any measurable feature c , let c_i^j be the value of c in document d_i^j . Let us develop the final formula in two steps.

Step 1. Define the stability of a feature in multiple versions of a single document.

Let S_i . Then the *stability* of a feature c in document d_i is defined to simply be the usual (normalized) entropy measure:

If c_i^j is a frequency measure, we can think of c_i^j / k_i as the probability that a random appearance of c in d_i is in version d_i^j . Then $H(\{c_i^j / k_i\})$ is just the usual entropy measure, normalized by $\log m$ to keep the range of stability values to $[0, 1]$. Thus, for example, if a feature assumed the identical value in every version of a document, its stability would be 1. If a feature assumed a positive value in a single version of the document but 0 in all others, its stability would be 0.

Step 2. Extend the definition to multiple documents $\{d_1, d_2, \dots, d_n\}$.

The impact that a given document has on the overall measure is defined to be proportional to the average value of the feature in the various versions of that document. (Thus, for example, when c is a frequency measure of some attribute, documents in which the attribute is more frequent will contribute proportionately more to the overall stability of the feature in the corpus.)

Let S . Then

This formula can be transformed to the equivalent formula:

(1)

4. Measuring Feature Stability Empirically Using Machine Translation

In order to empirically measure stability, one needs several text variants that convey the same meaning, or at least have substantial overlap. For example, we might consider translations of the same text by different translators, and to a much lesser extent, reports on the same event by different reporters or journalists. All these are relatively hard to obtain for experimental purposes. One interesting way to generate text variants artificially is to use a machine translation (MT) system to translate the text to another language and then translate it back to the original language. For our experiments, we used SystranPro to translate each document in the Reuters 21578 corpus into each of five different languages (French, German, Spanish, Italian, Portuguese) and back into English. In order to check that the measure is not overly dependent on the base corpus, we did the same to a set of several hundred book length documents from the British National Corpus. We applied formula (1) to frequencies of words, parts-of-speech n-grams and other linguistic features.

An obvious weakness of this experiment is that it is subject to the idiosyncrasies of Systran and, to a lesser extent, of the particular corpus. We have attempted to mitigate this affect by using five translation packages. This is only a partial solution since it is likely that all of them share certain underlying methods and programming code. Yet, as the experiments below indicate, this setting does provide an interesting “grasp” of feature stability, probably because the vast knowledge encoded in the five translation systems does capture much of the inherent phenomena that determine linguistic stability.

We will consider the stability distributions of various classes of features. We will also consider some specific features and discuss why their respective stabilities are particularly high or low.

5. Stability Distributions and Examples

In Figure 1, we show a histogram of stabilities of all single words in the Reuters corpus. As is evident the number of words in a given stability range descends as the mean of the range descends. When we look at specific word classes a clearer picture emerges. As might be expected, certain features, such as proper names, are highly stable. All proper names

have stability close to 1. Similarly, numbers have very high stability. Words with common synonyms such as *aid* (.37) and *help* (.82) are less stable, with the more common synonym more stable than the less common one. In extreme cases such as *huge* and *ratio*, stability is reduced to 0; the translation system always replaces them by more common synonyms such as *large* and *proportion*, respectively.

FIGURE 1 ABOUT HERE

Figure 1. Stability histogram of all words. The x-axis denotes stability values ranges and the y-axis denotes the proportion of features receiving the corresponding stability values.

In Figure 2(a-c), we show histograms of stability of nouns, verbs and function words, respectively. While nouns follow the general pattern with a plurality in the highest stability range, verbs and function words distribute more normally. This is because verbs are on average much more ambiguous than nouns. Similarly, most function words are also highly ambiguous and are often replaceable with equivalent syntactic constructions.

FIGURE 2 ABOUT HERE

Figure 2(a). Stability histogram for nouns

Figure 2(b). Stability histogram for verbs

Figure 2(c). Stability histogram for function words

In Table 1, we consider a selection of function words and their respective stabilities, using Reuters and BNC, respectively. A number of these examples are very instructive. Words like *and* and *the* don't offer more natural alternatives and are thus stable. However, words like *has* and *been* are unstable because, for example, the present perfect *has been* is easily replaced by the past tense *was*. Similarly, a word like *over* is easily replaced either by synonyms (e.g. *above*) or alternative constructions (e.g. *go over there* : *go there* : *go to there*). Note that with a few exceptions the stabilities yielded by Reuters and by BNC are quite close to each other.

TABLE 1 ABOUT HERE

Table 1. Examples of function words ranked by their respective stabilities using Reuters and BNC, respectively.

TABLE 2 ABOUT HERE

Table 2. Examples of noun-related part-of-speech triples ranked by their respective stabilities using Reuters and BNC, respectively.

Let's now consider features other than words. In Table 2, we consider a selection of trigrams of parts-of-speech and their respective stabilities. Note, for example, that the trigram *noun_noun_noun* is very unstable. A typical occurrence is "U.S. construction spending", a tightly-wound phrase invariably unwound into something looser like "spending on construction by the U.S."

Altogether, it can be seen that the stability measure captures in a unified way different types of semantic "uniqueness" that are related to both syntactic and lexical phenomena, including content words, function words and part-of-speech sequences.

6. Style-Based Text Categorization

Our main hypothesis is that frequent but unstable features are most useful for identifying an author's style. To assess this we will use frequency and stability as criteria for feature selection in text categorization experiments. First we give the necessary background to the text categorization and feature selection literature.

Style-based text categorization tasks, such as authorship attribution [Mosteller and Wallace, 1964; Holmes, 1995], are in a sense orthogonal to the more common problem of categorization by topic [Lewis and Ringuette, 1994; Schutze et al., 1995; Sebastiani, 2002]]. For style-based categorization, we seek features that are roughly invariant within the documents of a given author (or, more broadly, style class) but variant from author to author. Typically, a promising set of discriminating features is chosen and then training examples from each category and some machine-learning algorithm

are employed to produce a model for categorizing.

Since the number of potential discriminating features is often uncomfortably large, feature selection methods are sometimes used to select out a particularly promising set of features. A number of these methods, conveniently summarized in [Sebastiani, 2002], have proved to be reasonably successful for topical categorization. Typically, the features selected by these methods are those that, individually, discriminate well on the training corpus. While such methods do tend to eliminate useless features, they sometimes do harm by pre-empting the learning algorithms they are meant to serve: the learning algorithms themselves, by taking into account dependencies among features, eliminate useless features in more subtle ways than these direct feature selection methods. Furthermore, in the case of style-based categorization, these methods can lead to over-fitting by focusing attention on features that are highly correlated with one of the authors for reasons that might be specific for the topics discussed in the training corpus but unrelated to the author's generic style.

A different approach to feature selection for style-based categorization is that used by Mosteller and Wallace [1964] in their seminal work on authorship discrimination. Mosteller and Wallace [1964] simply chose a set of features that are not dependent on the training corpus but rather have certain appropriate universal properties. In their case, the features chosen were function words, which were deemed topic-independent. This is a perfectly sensible approach but it is somewhat limiting: it does not offer the flexibility of ranking features so that more or fewer can be chosen and it limits consideration in advance to a specific set of lexical features.

The main hypothesis of this paper is that those linguistic features that are both unstable and frequent are those that are most useful for style-based text categorization. Indeed, many function words are prime examples of frequent but unstable features; there are many other such features. The hypothesis makes intuitive sense: stable features are ones which don't offer viable meaning-preserving alternatives so that differences in usage of stable features between authors more likely reflect irrelevant differences in content than differences in style; differences in usage of frequent unstable features are likely to reflect different stylistic choices. Since instability can be ranked, we can choose more or fewer features as is needed, possibly using cross-validation to optimize. Moreover, as we have seen, unlike function words, the stability criterion can be applied to any type of linguistic feature, whether lexical, syntactic, complexity-based, etc.

A ranked list of the fifty most highly-ranked words according to the product of instability and log-frequency (using the Reuters and BNC corpora) is shown in the Appendix. Not that due to the specialized nature of the Reuter corpus, many financial words, which are not particularly relevant to other contexts, are highly ranked. Overall, of the 400 top words in

each ranked list (BNC and Reuters), 43% are common to both. We will see below that even when stability and frequency are measured using an “inappropriate” corpus, they yield superior results in the selection of potential style markers.

7. Experiments

In our first experiment, we will attempt to learn to distinguish the writing style of Anne Bronte from that of her sister, Charlotte Bronte. This is a particularly difficult attribution problem because the authors came from identical social and linguistic backgrounds and wrote in what appear to be very similar styles. We consider two books by Anne Bronte (*Agnes Gray*, *The Tenant of Wildfell Hall*) and two by Charlotte Bronte (*The Professor*, *Jane Eyre*). Each book is divided into between 100 and 150 equal-sized passages. We train on passages of one book by each author and test on passages of the remaining two books. We then run the experiment again with the training sets and test sets reversed and average the results. In the second experiment, we use a corpus of 260 fiction documents from the BNC, evenly split among male and female authors. We will attempt to learn the gender of a document's author [Koppel *et al.* 2002]. We test the effectiveness of our learning methods using five-fold cross-validation.

We begin with a feature set consisting of all features and then eliminate more and more features according to various criteria. For each reduced feature set we will use a learning algorithm to build a categorization model and then test the model on the chapters in the test set. We use Balanced Winnow [Littlestone, 1988; Dagan *et al.*, 1997] as our learning method.

In Figure 3(a), we show results on the Bronte experiment using Balanced Winnow on an initial feature set consisting of all 3500+ words that appear at least 4 times in the Bronte corpus. Feature reduction is performed by ranking features according to various measures (listed below). Since only 1300 words appear both in this list and in the Reuters corpus (which was used for measuring feature stability) the first data point we show for a reduced feature set is 1300. In Figure 3(b), we repeat the experiment using an initial feature set consisting of the full word list as well as all parts-of-speech triples that appear at least 5 times in the Bronte corpus. Each of the reduced sets consists of an equal number of words and parts-of-speech triples.

FIGURE 3 ABOUT HERE

Figure 3(a). Categorization accuracy of Balanced Winnow on the Bronte corpus using five feature reduction methods to select single words. The x-axis represents the number of features used and the y-axis records accuracy.

Figure 3(b). Categorization accuracy of Balanced Winnow on the Bronte corpus using five feature reduction methods to select single words and POS triples. The x-axis represents the number of features used and the y-axis records accuracy.

As our benchmark for the effectiveness of feature selection, we use the *odds ratio* (OR) measure [Mladenic, 1998; Ruiz and Srinivasan, 2002; Caropreso et al., 2002] which is a typical and particularly successful representative of discrimination-based feature reduction [Sebastiani 2002]. For a given feature c , let c_j be the frequency of c in category j and let $c_{j'}$ be the frequency of c in all other categories. Then OR ranks features according to the score . Other methods rearrange the same basic ingredients in different ways.

Altogether, five measures were used for ranking features:

OR – odds ratio in training corpus

F_t – average frequency in training corpus

$OR * F_t$ – odds ratio in training corpus * average frequency in training corpus

IN – instability ($= 1 - S(c)$)

$IN * F_r$ – instability * average frequency in Reuters

In the case of $IN * F_r$, we combine instability with frequency in Reuters rather than with frequency in the training corpus itself, in order to highlight the fact that the measure can be entirely corpus independent. The curve for $IN * F_t$ (not

shown) is very similar to that of IN^*F_r . Note also that, in order to emphasize the generality of the method, we measure both instability and frequency using the Reuters corpus rather than the BNC, which is more similar to the training corpus.

In Figures 4(a) and (b), we show the analogous results for five-fold cross-validation experiments on the gender experiment.

FIGURE 4 ABOUT HERE

Figure 4(a). Categorization accuracy of Balanced Winnow on the gender problem in BNC using five feature reduction methods to select single words. The x-axis represents the number of features used and the y-axis records accuracy.

Figure 4(b). Categorization accuracy of Balanced Winnow on the gender problem in BNC using five feature reduction methods to select single words and POS triples. The x-axis represents the number of features used and the y-axis records accuracy.

Note that, in both experiments, without taking feature frequency into account both *OR* and *IN* fail miserably: too many rare features are ranked highly by each method. More importantly, these experiments show that, although IN^*F_r is completely independent of the training corpus, it is actually the best measure by which to choose features. It is also evident that (a) feature selection does lead to improvement over using the complete feature set (i.e. letting Winnow select features implicitly); (b) optimal performance is maintained with a quite small feature set. In fact, in both experiments, the best 400 features selected according to this criterion are better than using a standard list of 400 function words, which achieves 81% accuracy on Bronte and 72% on the gender problem.

Note also that when the number of features approaches the bottom of our range, it is better to use the frequency of a feature in the training corpus than in Reuters. This is because many of the highly-ranked features based on Reuters frequency do not appear sufficiently frequently in the training/testing corpora to yield optimal categorization. This suggests that a broader-based corpus than Reuters is advisable.

Overall, these results suggest that the set of features identified as useful by IN^*F_r may constitute, to a substantial degree, a compact *universal* feature set for style-based categorization. This set might be thought of as a generalization of a universal feature set like a list of function words, with the added advantages that the features can be of any type (not

only lexical) and that they can be ranked.

8. Conclusions and Future Work

We have shown how the extent to which a linguistic feature may be replaced with a semantically equivalent feature can be effectively quantified. The resulting stability measure is useful for identifying promising candidates for style-based text categorization. The specific technique that we used for estimating stability – back and forth translation – is, admittedly, flawed due to its dependence on a particular software package but is, nevertheless, novel and effective. The results of our experiments indicate that we might regard frequent and unstable features as a generalization of function word lists that can serve as a universal feature set for style-based text categorization. More experiments – on other corpora, on more feature types, on other learning algorithms and using other existing feature reduction methods as benchmarks – can be performed to strengthen and extend these conclusions.

References

- [Barzilay and McKeown, 2001] Barzilay, R. and McKeown, K. 2001. Extracting Paraphrases from a Parallel Corpus. Proceedings of the Annual Meeting of the Association for Computational Linguistics (pp. 50-57).
- [Caropreso *et al.*, 2001] Caropreso, M. F., Matwin, S. and Sebastiani, F. 2001. A learner-independent evaluation of the usefulness of statistical phrases for automated text categorization. In A. G. Chin Ed., *Text Databases and Document Management: Theory and Practice*, pp. 78–102. Hershey, US: Idea Group Publishing.
- [Dagan *et al.*, 1997] Dagan, I., Karov, Y. and Roth, D. 1997. Mistake-Driven Learning in Text Categorization. In *Proceeding of Second Conference on Empirical Methods in Natural Language Processing (EMNLP-2)*, Providence, Rhode Island.
- [Holmes, 1995] Holmes, D. 1995. Authorship attribution. *Computers and the Humanities*, 28.
- [Koppel *et al.*, 2002] Koppel, M., Argamon S. and Shimoni, A. R. 2003. Automatically Categorizing Written Texts by Author Gender. *Literary and Linguistic Computing* 17(4), 401-412.
- [Lewis, 1997] Lewis, D. 1997. *Reuters-21578 text categorization test collection*.

- [Lewis and Ringuette, 1994] Lewis, D. and Ringuette, M. 1994. A Comparison of Two Learning Algorithms for Text Categorization. *Proceedings of the Third Annual Symposium on Document Analysis and Information Retrieval* (pp. 81-93), Las Vegas, USA.
- [Littlestone, 1988] Littlestone, N. 1988. Learning quickly when irrelevant features abound: A new linear-threshold algorithm. *Machine Learning*, 2.
- [Mladenic, 1998] Mladenic, D. 1998. Feature subset selection in text learning. In *Proceedings of ECML-98, 10th European Conference on Machine Learning* (pp. 95–100).
- [Mosteller and Wallace, 1964] Mosteller, F. and Wallace, D. L. 1964. *Applied Bayesian and Classical Inference: The Case of the Federalist Papers*. Springer.
- [Ruiz and Srinivasan, 2002] Ruiz, M. and Srinivasan, P. 2002. Hierarchical text classification using neural networks. *Information Retrieval* 5, 1, 87–118.
- [Schutze *et al.*, 1995] Schutze, H., Hull, D. A. and Pedersen, J. O. 1995. A comparison of classifiers and document representations for the routing problem. In *Proceedings of SIGIR-95, 18th ACM International Conference on Research and Development in Information Retrieval* (Seattle, US, 1995), pp. 229–237.
- [Sebastiani, 2002] Sebastiani, F. 2002. Machine learning in automated text categorization. *ACM Computing Surveys*, 34(1), 2002.
- [Shinyama *et al.*, 2002] Shinyama, Y., Sekine, S., Sudo, K. and Grishman, R. 2002. Automatic Paraphrase Acquisition from News Articles. *Proceedings of Human Language Technology Conference*, San Diego, USA.
- [Yang and Pedersen, 1997] Yang, Y. and Pedersen, J. O. 1997. A comparative study on feature selection in text categorization. In *Proceedings of ICML-97, 14th International Conference on Machine Learning* (pp. 412–420), Nashville, USA.

Appendix

The top 50 words in Reuters ranked by their respective $\text{Instability} \cdot \log(\text{Freq})$ scores.

1. indicated (.89)
2. has (.78)
3. known (.72)
4. which (.66)
5. them (.66)
6. dollar (.62)
7. order (.60)
8. at (.53)
9. series (.50)
10. over (.49)
11. supply (.49)
12. about (.47)
13. up (.46)
14. out (.45)
15. portion (.45)
16. share (.44)
17. have (.42)
18. his (.42)
19. benefit (.42)
20. state (.41)
21. he (.40)
22. says (.40)
23. shares (.40)
24. into (.38)
25. above (.38)
26. approximately (.38)
27. from (.38)
28. ratio (.38)
29. by (.38)
30. had (.37)
31. been (.37)
32. commerce (.36)

33. back (.36)
34. trade (.36)
35. debit (.35)
36. told (.35)
37. part (.34)
38. near (.34)
39. like (.34)
40. present (.34)
41. as (.34)
42. stock (.34)
43. connection (.33)
44. action (.32)
45. being (.32)
46. network (.31)
47. request (.31)
48. main (.31)
49. they (.31)
50. parts (.31)

The top 50 words in the BNC ranked by their respective $\text{Instability} * \log(\text{Freq})$ scores.

1. has (1.23)
2. which (.70)
3. his (.59)
4. her (.55)
5. into (.54)
6. order (.54)
7. over (.50)
8. him (.48)
9. about (.47)
10. thus (.43)
11. their (.41)
12. she (.40)
13. out (.40)
14. straight (.40)
15. top (.37)
16. started (.37)
17. away (.37)
18. up (.36)

19. do (.35)
20. off (.35)
21. just (.34)
22. above (.34)
23. by (.34)
24. makes (.34)
25. large (.33)
26. back (.33)
27. towards (.33)
28. everything (.33)
29. at (.32)
30. got (.32)
31. itself (.32)
32. however (.32)
33. indicated (.32)
34. outside (.31)
35. from (.31)
36. approximately (.31)
37. make (.31)
38. return (.31)
39. set (.31)
40. any (.30)
41. again (.29)
42. area (.29)
43. hour (.29)
44. does (.29)
45. went (.29)
46. far (.28)
47. being (.28)
48. low (.28)
49. start (.28)
50. part (.28)