Feature Instability as a Criterion for Selecting Potential Style Marke

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 whicl language element, such as a word or a syntactic construct, is replaceable by semantically equivalent elemer This measure may be perceived as quantifying the degree of available "synonymy" for a language item. ' 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 auth seek features that are typically used variably by different authors but are used consistently by any given author least, by the author whose writing we wish to distinguish. Obviously, these markers will differ from author to auth

In this paper, we wish to identify the pool of potentially useful linguistic features from which markers might fruit chosen. To be precise, we do not seek features that distinguish a particular author but rather author-independent 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 a this would certainly be worth noting. After all, the word *bad* is generally a reasonable, and more common, altern *awful*. The fact that our author chooses to use the word *awful* frequently therefore likely reflects deliberate s choice that we might profitably exploit for identifying the author's writing. What about the word *touchdow* frequency of use of *touchdown* is a rather poor feature for identifying an author's style. There are actually two reasons that this is so. First, there is no alternative word for expressing the concept *touchdown* so that its use d reflect stylistic choice. Second, *touchdown* is tightly tied to a particular topic so that its frequency of use in one co unlikely to reflect the frequency with which it will be used in other documents, which might concern other top contrast, *awful* is commonly used across most topics and has plausible alternatives. Some words satisfy one of th criteria but not both. For example, *perspire* offers a plausible alternative (*sweat*) but is not frequently used across range of topics. On the other hand, words like *blue* and *ten* are used across topics but don't have common alternati

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

2. Linguistic Feature Stability

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

In this paper, we introduce a new type of measure for linguistic elements that we call *meaning-preserving stab* *stability* for short. It captures the degree to which a linguistic element or construct can be substituted for in a piece without affecting the text meaning. This measure is applied most intuitively to words and terms, but is applicable other types of linguistic constructs, such as part-of-speech sequences. Stability captures interesting properties in la and thus seems interesting from a purely scientific point of view, but it also has potential use within applie Stability can be perceived as quantifying the typical degree of available "synonymy" for a language element generalizing the notion of synonymy to any type of linguistic feature rather than being restricted to its common 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 v (*John, window*), while others are replaced by other words (*was : had been, couch : sofa, next to : by : near,* *reclining*) which don't significantly change the meaning of the sentence. More generally, consider features of a te as word or phrase frequencies, frequencies of syntactic structures or any other feature whose value can be measu given text. Roughly speaking, the *stability* of such a feature is the extent to which the measured value of that tends to remain invariant across different texts that convey the same meaning. For example, proper nouns are very while words with many synonyms are unstable.

Obtaining training material for measuring stability is a difficult challenge, requiring alternative versions of te carry – in the ideal case, exactly – the same meaning. "Parallel" (monolingual) texts of this nature have been u automatic paraphrase extraction, using either multiple translations of the same text or news stories from different that describe the same event [Barzilay and McKeown 2001; Shinyama et al. 2002]. The disadvantage of these t training material is that they rely on the fact that different people have manually created different versions of th text. Such manually created versions are not easily available for most texts. In order to lighten the supe requirements, we have used a machine translation (MT) system to translate our training corpus back-and-fo multiple languages, thus obtaining several English versions of the same text. Indeed, the use of automatic transla generate training data is problematic for several reasons, as discussed below. Yet, the combination of MT wit technologies that can leverage the stability measure, such as style-based text categorization, provides ap 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 measu then present empirical measurements of stability and illustrate the properties that it captures. Finally, we demo 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 different meaning-preserving variants of a text. We formally define a specific stability measure and consider en results of stability experiments on the Reuters 21578 corpus [Lewis, 1997] and on the British National Corpus (BN

Let $\{d_1,d_2,\ldots,d_n\}$ be a set of documents (or text segments) and let $\{d_i^1,d_i^2,\ldots,d_i^m\}$ be $m > 1$ different versions of *a* the meanings of the m versions are all roughly identical.[1] For any measurable feature *c*, let c_i^j be the value 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 $k_i = \Sigma_j\, c_i^j$. Then the *stability* of a feature *c* in document d_i is defined to simply be the usual (normalized) measure:

$$H(c_i) = -\Sigma_j\, [(c_i^j/\, k_i)\log\, (c_i^j/\, k_i)]]/\log m$$

[1] Since stability is an empirical measure based on statistical data, its definition does not depend on the exact notion of "r preservation" that holds across the different text versions. Applying the stability measure would make sense whenever the text versions represent alternative interchangeable linguistic choices.

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 ver
Then $H(c_i)$ is just the usual entropy measure, normalized by log m to keep the range of stability values to [0,1]. T
example, if a feature assumed the identical value in every version of a document, its stability would be 1. If a
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 $\{d_1, d_2, \ldots, d_n\}$.

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

Let $K = \Sigma_i k_i$. Then

$$S(c) = \Sigma_i [(k_i/K) * H(c_i)]$$

This formula can be transformed to the equivalent formula:

$$S(c) = \{\Sigma_i [k_i \log k_i - \Sigma_j c_i^j \log c_i^j]\} / K*\log m \quad (1)$$

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

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

We will consider the stability distributions of various classes of features. We will also consider some specific f
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.

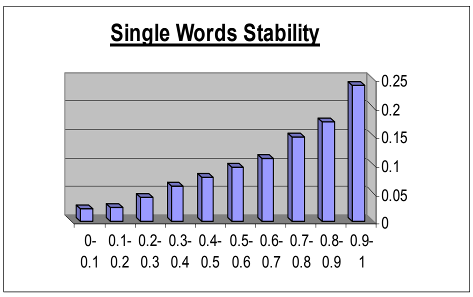


Figure 1. Stability histogram of all words. The x-axis denotes stability value ranges and the y-axis denotes the proportion of features receiving the corresponding stability values.

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

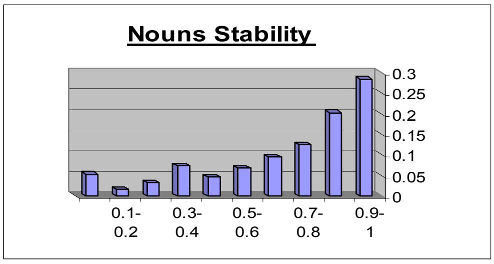


Figure 2(a). Stability histogram for nouns

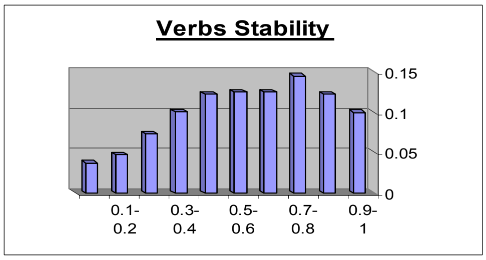


Figure 2(b). Stability histogram for verbs

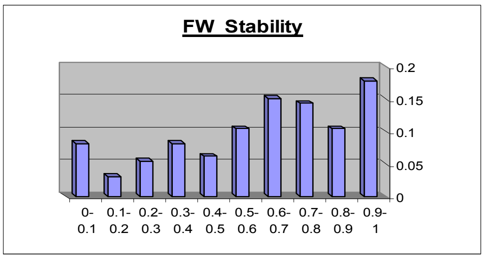


Figure 2(c): Stability histogram for function words

function words are also highly ambiguous and are often replaceable with equivalent syntactic constructions.

FW	**Reuters**	**BNC**	**FW**	**Reuters**	**l**
and	0.99	0.99	from	0.77	(
the	0.97	0.99	by	0.76	(
not	0.95	0.97	been	0.66	(
is	0.94	0.95	which	0.56	(
of	0.94	0.95	has	0.52	(
it's	0.87	0.88	over	0.47	(

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

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

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

Altogether, it can be seen that the stability measure captures in a unified way different types of semantic "uniqı that are related to both syntactic and lexical phenomena, including content words, function words and part-of 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 ass we will use frequency and stability as criteria for feature selection in text categorization experiments. First we ￼ 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] a sense orthogonal to the more common problem of categorization by topic [Lewis and Ringuette, 1994; Schutz 1995; Sebastiani, 2002]]. For style-based categorization, we seek features that are roughly invariant wit 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.

POS	Reuters Stability	BN Stabi
NN_IN_DT	0.94	0.97
IN_DT_NN	0.94	0.98
DT_NN_IN	0.93	0.97
NN_IN_NN	0.92	0.89
IN_NN_IN	0.86	0.86
NN_IN_NNP	0.81	0.89
IN_CD_NN	0.76	0.72
CD_NN_IN	0.72	0.58
JJ_NN_NN	0.67	0.80
NN_NN_NN	0.43	0.59

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

The main hypothesis of this paper is that those linguistic features that are both unstable and frequent are those most useful for style-based text categorization. Indeed, many function words are prime examples of frequent but u features; there are many other such features. The hypothesis makes intuitive sense: stable features are ones whic offer viable meaning-preserving alternatives so that differences in usage of stable features between authors mor reflect irrelevant differences in content than differences in style; differences in usage of frequent unstable featu likely to reflect different stylistic choices. Since instability can be ranked, we can choose more or fewer featur needed, possibly using cross-validation to optimize. Moreover, as we have seen, unlike function words, the s 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 (us Reuters and BNC corpora) is shown in the Appendix. Not that due to the specialized nature of the Reuter corpus financial words, which are not particularly relevant to other contexts, are highly ranked. Overall, of the 400 top w each ranked list (BNC and Reuters), 43% are common to both. We will see below that even when stability and fre are measured using an "inappropriate" corpus, they yield superior results in the selection of potential style marker

7. Experiments

In our first experiment, we will attempt to learn to distinguish the writing style of Anne Bronte from that of he Charlotte Bronte. This is a particularly difficult attribution problem because the authors came from identical soc linguistic backgrounds and wrote in what appear to be very similar styles. We consider two books by Anne (*Agnes Gray, The Tenant of Wildfell Hall*) and two by Charlotte Bronte (*The Professor, Jane Eyre*). Each book is into between 100 and 150 equal-sized passages. We train on passages of one book by each author and test on pass the remaining two books. We then run the experiment again with the training setsand test sets reversed and aver results. In the second experiment, we use a corpus of 260 fiction documents from the BNC, evenly split among m female authors. We will attempt to learn the gender of a document's author [Koppel *et al.* 2002]. We t 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 criteria. For each reduced feature set we will use a learning algorithm to build a categorization model and then model on the chapters in the test set. We use Balanced Winnow [Littlestone, 1988; Dagan et al., 1997] as our l method.

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

FIGURE 3 ABOUT HERE

As our benchmark for the effectiveness of feature selection, we use the *odds ratio* (OR) measure [Mladenic Ruiz and Srinivasan, 2002; Caropreso et al., 2002] which is a typical and particularly successful represent discrimination-based feature reduction [Sebastiani 2002]. For a given feature c, let c_j be the frequency of c in ca and let $c_{j'}$ be the frequency of c in all other categories[2]. Then OR ranks features according to the score $\Sigma_j\, c_j(1\text{–}c_j)/$ 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 itself, in order to highlight the fact that the measure can be entirely corpus independent. The curve for $IN^*\, F_t$ (not is very similar to that of $IN^*\, F_r$. Note also that, in order to emphasize the generality of the method, we measu instability and frequency using the Reuters corpus rather than the BNC, which is more similar to the training corp

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

FIGURE 4 ABOUT HERE

Note that, in both experiments, without taking feature frequency into account both *OR* and *IN* fail miserably: to rare features are ranked highly by each method. More importantly, these experiments show that, although *I* completely independent of the training corpus, it is actually the best measure by which to choose features. It evident that (a) feature selection does lead to improvement over using the complete feature set (i.e. letting Winno features implicitly); (b) optimal performance is maintained with a quite small feature set. In fact, in both experime best 400 features selected according to this criterion are better than using a standard list of 400 function words 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 frequen feature in the training corpus than in Reuters. This is because many of the highly-ranked features based on

---

[2] Technically, the frequency used in *OR* as defined by Caropreso et al (2002) is the proportion of documents in the category i the feature appears at least once. Among common features, which appear in every document, *OR* as so defined o discrimination. Thus it is inappropriate for style-based categorization. Instead, we use the actual relative frequency of the featu category.

frequency do not appear sufficiently frequently in the training/testing corpora to yield optimal categorizatio 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 sub degree, a compact *universal* feature set for style-based categorization. This set might be thought of as a generaliz a universal feature set like a list of function words, with the added advantages that the features can be of any ty 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 feat be effectively quantified. The resulting stability measure is useful for identifying promising candidates for styl text categorization. The specific technique that we used for estimating stability – back and forth translatio admittedly, flawed due to its dependence on a particular software package but is, nevertheless, novel and effecti results of our experiments indicate that we might regard frequent and unstable features as a generalization of f word lists that can serve as a universal feature set for style-based text categorization. More experiments – o corpora, on more feature types, on other learning algorithms and using other existing feature reduction metl 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 (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 evaluatio usefulness of statistical phrases for automated text categorization. In A. G. Chin Ed., *Text Databases and Do 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 Categorizati *Proceeding of Second Conference on Empirical Methods in Natural Language Processing (EMNLP-2)*, Prov 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 T 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 f Categorization. *Proceedings of the Third Annual Symposium on Document Analysis and Information Retrieval* (93), Las Vegas, USA.

[Littlestone, 1988] Littlestone, N. 1988. Learning quickly when irrelevant features abound: A new linear-th algorithm. *Machine Learning*, 2.

[Mladenic, 1998] Mladenic, D. 1998. Feature subset selection in text learning. In *Proceedings of ECML-9 European Conference on Machine Learning* (pp. 95–100).

[Mosteller and Wallace, 1964] Mosteller, F. and Wallace, D. L. 1964. *Applied Bayesian and Classical Inferen Case of the Federalist Papers*. Springer.

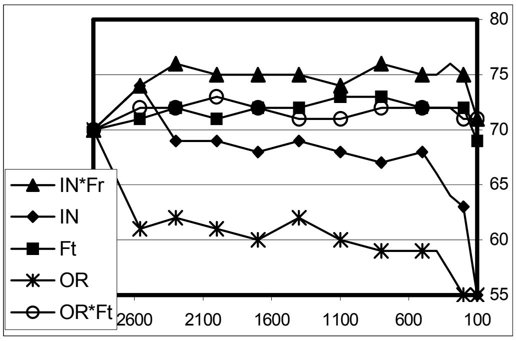
[Ruiz and Srinivasan, 2002] Ruiz, M. and Srinivasan, P. 2002. Hierarchical text classification using neural ne *Information Retrieval* 5, 1, 87–118.

[Schutze *et al.*, 1995] Schutze, H., Hull, D. A. and Pedersen, J. O. 1995. A comparison of classifiers and do representations for the routing problem. In *Proceedings of SIGIR-95, 18th ACM International Conference on R and Development in Information Retrieval* (Seattle, US, 1995), pp. 229–237.

[Sebastiani, 2002] Sebastiani, F. 2002. Machine learning in automated text categorization**.** *ACM Computing S* 34(1), 2002.

[Shinyama *et al*., 2002] Shinyama, Y., Sekine, S., Sudo, K. and Grishman, R. 2002. Automatic Paraphrase Acq 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 categorization. In *Proceedings of ICML-97, 14th International Conference on Machine Learning* (pp. 41 Nashville, USA.



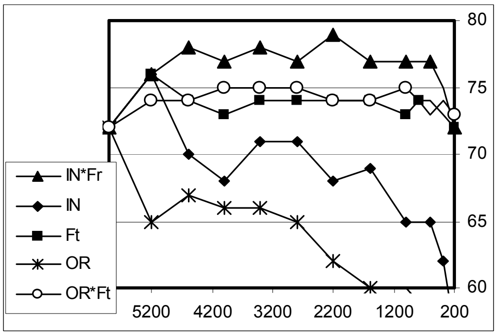
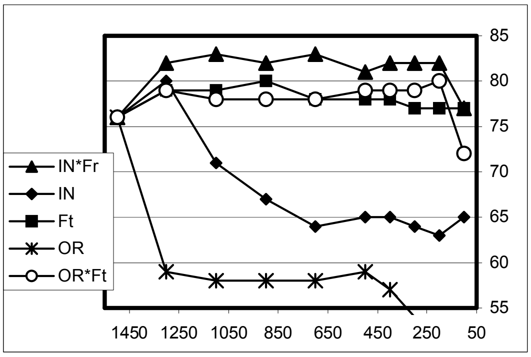


Figure4(a). Categorization accuracy of Balanced Winnow on the gender problem in BNC using five feature reduction methods to select single words. The x-axis represents the number of features used and the y-axis records accuracy.

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



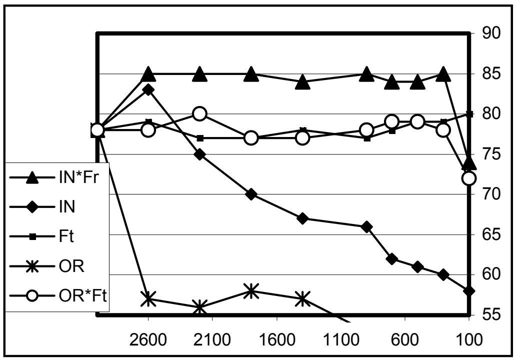


Figure 3(a). Categorization accuracy of Balanced Winnow on the Bronte corpus using five feature reduction methods to select single words. The x-axis represents the number of features used and the y-axis records accuracy.

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.

Appendix

The top 50 words in Reuters ranked by their respective Instability*log(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
41. as
42. stock
43. connecti
44. action
45. being
46. network
47. request
48. main
49. they
50. parts

The top 50 words in the BNC ranked by their respective Instability*log(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
41. again
42. area
43. hour
44. does
45. went
46. far
47. being
48. low
49. start
50. part