Temporal Context: Applications and Implications for Computational Linguistics

Robert A. Liebscher
Department of Cognitive Science
University of California, San Diego
La Jolla, CA 92037
rliebsch@cogsci.ucsd.edu

Abstract

This paper describes several ongoing projects that are united by the theme of changes in lexical use over time. We show that paying attention to a document's temporal context can lead to improvements in information retrieval and text categorization. We also explore a potential application in document clustering that is based upon different types of lexical changes.

1 Introduction

Tasks in computational linguistics (CL) normally focus on the *content* of a document while paying little attention to the *context* in which it was produced. The work described in this paper considers the importance of temporal context. We show that knowing one small piece of information–a document's publication date–can be beneficial for a variety of CL tasks, some familiar and some novel.

The field of historical linguistics attempts to categorize changes at all levels of language use, typically relying on data that span centuries (Hock, 1991). The recent availability of very large textual corpora allows for the examination of changes that take place across shorter time periods. In particular, we focus on lexical change across decades in corpora of academic publications and show that the changes can be fairly dramatic during a relatively short period of time.

As a preview, consider Table 1, which lists the top five unigrams that best distinguished the field of computational linguistics at different points in time, as derived from the ACL proceedings[1] using the odds ratio measure (see Section 3). One can quickly glean that the field has become increasingly empirical through time.

1979-84	1985-90	1991-96	1997-02
system	phrase	discourse	word
natural	plan	tree	corpus
language	structure	algorithm	training
knowledge	logical	unification	model
database	interpret	plan	data

Table 1: ACL's most characteristic terms for four time periods, as measured by the odds ratio

With respect to academic publications, the very nature of the enterprise forces the language used within a discipline to change. An author's word choice is shaped by the preceding literature, as she must say something novel while placing her contribution in the context of what has already been said. This begets neologisms, new word senses, and other types of changes.

This paper is organized as follows: In Section 2, we introduce *temporal term weighting*, a technique that implicitly encodes time into keyword weights to enhance information retrieval. Section 3 describes the technique of *temporal feature modification*, which exploits temporal information to improve the text categorization task. Section 4 introduces several types of lexical changes and a potential application in document clustering.

[1]The details of each corpus used in this paper can be found in the appendix.

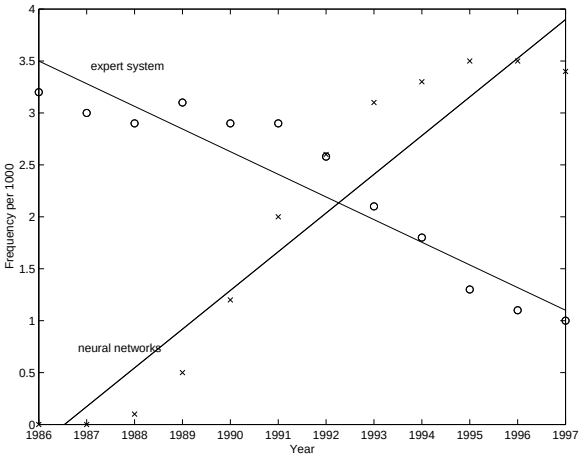

Figure 1: Changing frequencies in AI abstracts

2 Time in information retrieval

In the task of retrieving relevant documents based upon keyword queries, it is customary to treat each document as a vector of terms with associated "weights". One notion of term weight simply counts the occurrences of each term. Of more utility is the scheme known as *term frequency-inverse document frequency* (TF.IDF):

$$w_{kd} = f_{kd} \cdot log(N/D_k)$$

where w_{kd} is the weight of term k in document *d*, f_{kd} is the frequency of k in *d*, *N* is the total number of documents in the corpus, and D_k is the total number of documents containing k. Very frequent terms (such as function words) that occur in many documents are downweighted, while those that are fairly unique have their weights boosted.

Many variations of TF.IDF have been suggested (Singhal, 1997). Our variation, *temporal term weighting* (TTW), incorporates a term's IDF at different points in time:

$$w_{kd} = f_{kd} \cdot log(N(t)/D_k(t))$$

Under this scheme, the document collection is divided into *T* time slices, and *N* and D_k are computed for each slice *t*. Figure 1 illustrates why such a modification is useful. It depicts the frequency of the terms `neural networks` and `expert system` for each year in a collection of Artificial Intelligence-related dissertation abstracts. Both terms follow a fairly linear trend, moving in opposite directions.

As was demonstrated for CL in Section 1, the terms which best characterize AI have also changed through time. Table 2 lists the top five "rising" and "falling" bigrams in this corpus, along with their least-squares fit to a linear trend. Lexical variants (such as plurals) are omitted. Using an *a*temporal TF.IDF, both rising and falling terms would be assigned weights proportional only to f_{kd}. A novice user issuing a query would be given a temporally random scattering of documents, some of which might be state-of-the-art, others very outdated.

But with TTW, the weights are proportional to the collective "community interest" in the term at a given point in time. In academic research documents, this yields two benefits. If a term rises from obscurity to popularity over the duration of a corpus, it is not unreasonable to assume that this term originated in one or a few *seminal* articles. The term is not very frequent across documents when these articles are published, so its weight in the seminal articles will be amplified. Similarly, the term will be downweighted in articles when it has become ubiquitous throughout the literature.

For a falling term, its weight in early documents will be dampened, while its later use will be emphasized. If a term is very frequent in a document after it has been relegated to obscurity, this is likely to be an *historical review* article. Such an article would be a good place to start an investigation for someone who is unfamiliar with the term.

Term	**r**
neural network	0.9283
fuzzy logic	0.9035
genetic algorithm	0.9624
real world	0.8509
reinforcement learning	0.8447
artificial intelligence	-0.9309
expert system	-0.9241
knowledge base	-0.9144
problem solving	-0.9490
knowledge representation	-0.9603

Table 2: Rising and falling AI terms, 1986-1997

2.1 Future work

We have discovered clear frequency trends over time in several corpora. Given this, TTW seems beneficial for use in information retrieval, but is in an embryonic stage. The next step will be the development and implementation of empirical tests.

IR systems typically are evaluated by measures such as precision and recall, but a different test is necessary to compare TTW to an atemporal TF.IDF. One idea we are exploring is to have a system explicitly tag seminal and historical review articles that are centered around a query term, and then compare the results with those generated by bibliometric methods. Few bibliometric analyses have gone beyond examinations of citation networks and the keywords associated with each article. We would consider the entire text.

3 Time in text categorization

Text categorization (TC) is the problem of assigning documents to one or more pre-defined categories. As Section 2 demonstrated, the terms which best characterize a category can change through time, so intelligent use of temporal context may prove useful in TC.

Consider the example of sorting newswire documents into the categories **ENTERTAINMENT**, **BUSINESS**, **SPORTS**, **POLITICS**, and **WEATHER**. Suppose we come across the term athens in a training document. We might expect a fairly uniform distribution of this term throughout the five categories; that is, $Pr(\mathbf{C}|\texttt{athens})$ = 0.20 for each **C**. However, in the summer of 2004, we would expect $Pr(\mathbf{SPORTS}|\texttt{athens})$ to be greatly increased relative to the other categories due to the city's hosting of the Olympic games.

Documents with "temporally perturbed" terms like athens contain potentially valuable information, but this is lost in a statistical analysis based purely on the content of each document, irrespective of its temporal context. This information can be recovered with a technique we call *temporal feature modification* (TFM). We first outline a formal model of its use.

Each term k is assumed to have a *generator* G^k that produces a "true" distribution $Pr(\mathbf{C}|\mathsf{k})$ across all categories. External events at time *y* can perturb k's generator, causing $Pr(\mathbf{C}|\mathsf{k})_y$ to be different relative to the background $Pr(\mathbf{C}|\mathsf{k})$ computed over the entire corpus. If the perturbation is significant, we want to separate the instances of k at time *y* from all other instances. We thus treat athens and "athens+summer2004" as though they were actually *different* terms, because they came from two different generators.

TFM is a two step process that is captured by this pseudocode:

```
VOCABULARY ADDITIONS:
for each class C:
  for each year y:
    PreModList(C,y,L) = OddsRatio(C,y,L)
    ModifyList(y) =
      DecisionRule(PreModList(C,y,L))
    for each term k in ModifyList(y):
      Add pseudo-term "k+y" to Vocab

DOCUMENT MODIFICATIONS:
for each document:
  y = year of doc
  for each term k:
    if "k+y" in Vocab:
      replace k with "k+y"
  classify modified document
```

PreModList(**C**,*y*,*L*) is a list of the top *L* lexemes that, by the odds ratio measure[2], are highly associated with category **C** in year *y*. We test the hypothesis that these come from a perturbed generator in year *y*, as opposed to the atemporal generator G^k, by comparing the odds ratios of term-category pairs in a *PreModList* in year *y* with the same pairs across the entire corpus. Terms which pass this test are added to the final *ModifyList*(*y*) for year *y*. For the results that we report, *DecisionRule* is a simple ratio test with threshold factor *f*. Suppose *f* is 2.0: if the odds ratio between **C** and k is twice as great in year *y* as it is atemporally, the decision rule is "passed". The generator $G^k$is considered perturbed in year *y* and k is added to *ModifyList*(*y*). In the training and testing phases, the documents are modified so that a term k is replaced with the pseudo-term "k+*y*" if it passed the ratio test.

3.1 ACM Classifications

We tested TFM on corpora representing genres from academic publications to Usenet postings,

[2] Odds ratio is defined as $p_k(1-q_k)/q_k(1-p_k)$, where *p* is *Pr*(k|**C**), the probability that term k is present given category **C**, and *q* is *Pr*(k|!**C**).

Corpus	Vocab size	No. docs	No. cats
SIGCHI	4542	1910	20
SIGPLAN	6744	3123	22
DAC	6311	2707	20

Table 3: Corpora characteristics. Terms occurring at least twice are included in the vocabulary.

and it improved classification accuracy in every case. The results reported here are for abstracts from the proceedings of several of the Association for Computing Machinery's conferences: SIGCHI, SIGPLAN, and DAC. TFM can benefit the ACM community through *retrospective categorization* in two ways: (1) 7.73% of abstracts (nearly 6000) across the entire ACM corpus that are expected to have category labels do not have them; (2) When a group of terms becomes popular enough to induce the formation of a new category, a frequent occurrence in the computing literature, TFM would separate the "old" uses from the "new" ones.

The ACM classifies its documents in a hierarchy of four levels; we used an aggregating procedure to "flatten" these. The characteristics of each corpus are described in Table 3. The "TC minutiae" used in these experiments are: Stoplist, Porter stemming, 90/10% train/test split, Laplacian smoothing. Parameters such as type of classifier (Naïve Bayes, KNN, TF.IDF, Probabilistic indexing) and threshold factor f were varied.

3.2 Results

Figure 2 shows the improvement in classification accuracy for different percentages of terms modified, using the best parameter combinations for each corpus, which are noted in Table 4. A baseline of 0.0 indicates accuracy without any temporal modifications. Despite the relative paucity of data in terms of document length, TFM still performs well on the abstracts. The actual accuracies when no terms are modified are less than stellar, ranging from 30.7% (DAC) to 33.7% (SIGPLAN) when averaged across all conditions, due to the difficulty of the task (20-22 categories; each document can only belong to one). Our aim is simply to show improvement.

In most cases, the technique performs best when

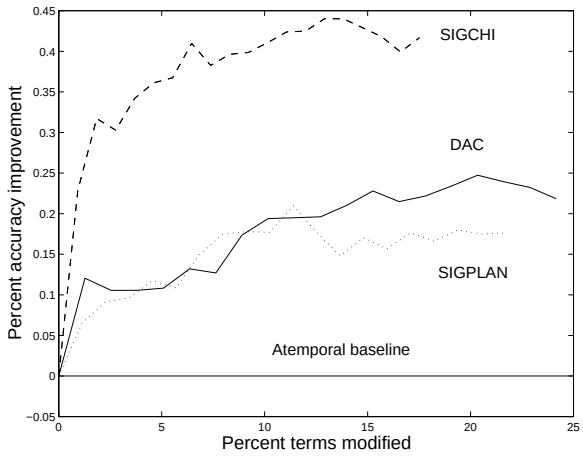

Figure 2: Improvement in categorization performance with TFM, using the best parameter combinations for each corpus

making relatively few modifications: the left side of Figure 2 shows a rapid performance increase, particularly for SIGCHI, followed by a period of diminishing returns as more terms are modified. After requiring the one-time computation of odds ratios in the training set for each category/year, TFM is very fast and requires negligible extra storage space.

3.3 Future work

The "bare bones" version of TFM presented here is intended as a proof-of-concept. Many of the parameters and procedures can be set arbitrarily. For initial feature selection, we used odds ratio because it exhibits good performance in TC (Mladenic, 1998), but it could be replaced by another method such as information gain. The ratio test is not a very sophisticated way to choose which terms should be modified, and presently only detects the *surges* in the use of a term, while ignoring the (admittedly rare) declines.

Using TFM on a Usenet corpus that was more balanced in terms of documents per category and per year, we found that allowing different terms to "compete" for modification was more effective than the egalitarian practice of choosing L terms from each category/year. There is no reason to believe that each category/year is equally likely to contribute temporally perturbed terms.

Finally, we would like to exploit temporal *con-*

Corpus	Improvement	Classifier	n-gram size	Vocab frequency min.	Ratio threshold f
SIGCHI	41.0%	TF.IDF	Bigram	10	1.0
SIGPLAN	19.4%	KNN	Unigram	10	1.0
DAC	23.3%	KNN	Unigram	2	1.0

Table 4: Top parameter combinations for TFM by improvement in classification accuracy. *Vocab frequency min.* is the minimum number of times a term must appear in the corpus in order to be included.

tiguity. The present implementation treats time slices as independent entities, which precludes the possibility of discovering temporal trends in the data. One way to incorporate trends *implicitly* is to run a smoothing filter across the temporally aligned frequencies. Also, we treat each slice at annual resolution. Initial tests show that aggregating two or more years into one slice improves performance for some corpora, particularly those with temporally sparse data such as DAC.

4 Future work

A third part of this research program, presently in the exploratory stage, concerns *lexical (semantic) change*, the broad class of phenomena in which words and phrases are coined or take on new meanings (Bauer, 1994; Jeffers and Lehiste, 1979). Below we describe an application in document clustering and point toward a theoretical framework for lexical change based upon recent advances in network analysis.

Consider a scenario in which a user queries a document database for the term `artificial intelligence`. We would like to create a system that will cluster the returned documents into three categories, corresponding to the types of change the query has undergone. These responses illustrate the three categories, which are not necessarily mutually exclusive:

1. "This term is now more commonly referred to as `AI` in this collection",

2. "These documents are *about* `artificial intelligence`, though it is now more commonly called `machine learning`",

3. "The following documents are *about* `artificial intelligence`, though in this collection its use has become *tacit*".

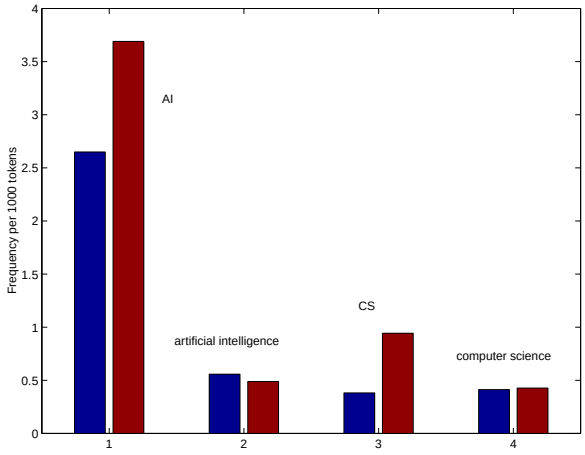

Figure 3: Frequencies in the first (left bar) and second (right bar) halves of an AI discussion forum

4.1 Acronym formation

In Section 2, we introduced the notions of "rising" and "falling" terms. Figure 3 shows relative frequencies of two common terms and their acronyms in the first and second halves of a corpus of AI discussion board postings collected from 1983-1988. While the acronyms increased in frequency, the expanded forms decreased or remained the same. A reasonable conjecture is that in this *informal* register, the acronyms `AI` and `CS` largely replaced the expansions. During the same time period, the more *formal* register of dissertation abstracts did not show this pattern for any acronym/expansion pairs.

4.2 Lexical replacement

Terms can be replaced by their acronyms, or by other terms. In Table 1, `database` was listed among the top five terms that were most characteristic of the ACL proceedings in 1979-1984. Bisecting this time slice and including bi-

grams in the analysis, `data base` ranks higher than `database` in 1979-1981, but drops much lower in 1982-1984. Within this brief period of time, we see a lexical replacement event taking hold. In the AI dissertation abstracts, `artificial intelligence` shows the greatest decline, while the conceptually similar terms `machine learning` and `pattern recognition` rank sixth and twelfth among the top rising terms.

There are social, geographic, and linguistic forces that influence lexical change. One example stood out as having an easily identified cause: political correctness. In a corpus of dissertation abstracts on communication disorders from 1982-2002, the term `subject` showed the greatest relative decrease in frequency, while `participant` showed the greatest increase. Among the top ten bigrams showing the sharpest declines were three terms that included the word `impaired` and two that included `disabled`.

4.3 "Tacit" vocabulary

Another, more subtle lexical change involves the gradual *disappearance* of terms due to their increasingly "tacit" nature within a particular community of discourse. Their existence becomes so obvious that they need not be mentioned within the community, but would be necessary for an outsider to fully understand the discourse.

Take, for example, the terms `backpropagation` and `hidden layer`. If a researcher of neural networks uses these terms in an abstract, then `neural network` does not even warrant printing, because they have come to *imply* the presence of `neural network` within this research community.

Applied to IR, one might call this "retrieval by implication". Discovering tacit terms is no simple matter, as many of them will not follow simple *is-a* relationships (e.g. `terrier` is a `dog`). The example of the previous paragraph seems to contain a hierarchical relation, but it is difficult to define. We believe that examining the temporal trajectories of closely related *networks* of terms may be of use here, and is also part of a more general project that we hope to undertake. Our intention is to improve existing models of lexical change using recent advances in network analysis (Barabasi et al., 2002; Dorogovtsev and Mendes, 2001).

References

A. Barabasi, H. Jeong, Z. Neda, A. Schubert, and T. Vicsek. 2002. Evolution of the social network of scientific collaborations. *Physica A*, 311:590–614.

L. Bauer. 1994. *Watching English Change*. Longman Press, London.

S. N. Dorogovtsev and J. F. F. Mendes. 2001. Language as an evolving word web. *Proceedings of The Royal Society of London, Series B*, 268(1485):2603–2606.

H. H. Hock. 1991. *Principles of Historical Lingusitics*. Mouton de Gruyter, Berlin.

R. J. Jeffers and I. Lehiste. 1979. *Principles and Methods for Historical Lingusitics*. The MIT Press, Cambridge, MA.

D. Mladenic. 1998. *Machine Learning on non-homogeneous, distributed text data*. Ph.D. thesis, University of Ljubljana, Slovenia.

A. Singhal. 1997. *Term weighting revisited*. Ph.D. thesis, Cornell University.

Appendix: Corpora

The corpora used in this paper, preceded by the section in which they were introduced:

1: The annual proceedings of the Association for Computational Linguistics conference (1978-2002). Accessible at http://acl.ldc.upenn.edu/.

2: Over 5000 PhD and Masters dissertation abstracts related to Artificial Intelligence, 1986-1997. Supplied by University Microfilms Inc.

3.1: Abstracts from the ACM-IEEE Design Automation Conference (DAC; 1964-2002), Special Interest Groups in Human Factors in Computing Systems (SIGCHI; 1982-2003) and Programming Languages (SIGPLAN; 1973-2003). Supplied by the ACM. See also Table 3.

3.3: Hand-collected corpus of six discussion groups: misc.consumers, alt.atheism, rec.arts.books, comp.{arch, graphics.algorithms, lang.c}. Each group contains 1000 documents per year from 1993-2002. Viewable at http://groups.google.com/.

4.2: Over 4000 PhD and Masters dissertation abstracts related to communication disorders, 1982-2002. Supplied by University Microfilms Inc.