Is Knowledge-Free Induction of Multiword Unit Dictionary Headwords a Solved Problem?

Patrick Schone and Daniel Jurafsky
University of Colorado, Boulder CO 80309
{schone, jurafsky}@cs.colorado.edu

Abstract

We seek a knowledge-free method for inducing multiword units from text corpora for use as machine-readable dictionary headwords. We provide two major evaluations of nine existing collocation-finders and illustrate the continuing need for improvement. We use Latent Semantic Analysis to make modest gains in performance, but we show the significant challenges encountered in trying this approach.

1 Introduction

A multiword unit (MWU) is a *connected* collocation: a sequence of neighboring words "whose exact and unambiguous meaning or connotation cannot be derived from the meaning or connotation of its components" (Choueka, 1988). In other words, MWUs are typically non-compositional at some linguistic level. For example, *phonological* non-compositionality has been observed (Finke & Weibel, 1997; Gregory, et al, 1999) where words like "got" [gɑt] and "to" [tu] change phonetically to "gotta" [gɑrə] when combined. We have interest in inducing headwords for machine-readable dictionaries (MRDs), so our interest is in *semantic* rather than phonological non-compositionality. As an example of semantic non-compositionality, consider "compact disk": one could not deduce that it was a music medium by only considering the semantics of "compact" and "disk."

MWUs may also be non-substitutable and/or non-modifiable (Manning and Schütze, 1999). Non-substitutability implies that substituting a word of the MWU with its synonym should no longer convey the same original content: "compact disk" does not readily imply "densely-packed disk." Non-modifiability, on the other hand, suggests one cannot modify the MWU's structure and still convey the same content: "compact disk" does not signify "disk that is compact."

MWU dictionary headwords generally satisfy at least one of these constraints. For example, a compositional phrase would typically be excluded from a hard-copy dictionary since its constituent words would already be listed. These strategies allow hard-copy dictionaries to remain compact.

As mentioned, we wish to find MWU headwords for machine-readable dictionaries (MRDs). Although space is not an issue in MRDs, we desire to follow the lexicographic practice of reducing redundancy. As Sproat indicated, "simply expanding the dictionary to encompass every word one is ever likely to encounter is wrong: it fails to take advantage of regularities" (1992, p. xiii). Our goal is to identify an automatic, knowledge-free algorithm that finds all and only those collocations where it *is* necessary to supply a definition. "Knowledge-free" means that the process should proceed without human input (other than, perhaps, indicating whitespace and punctuation).

This seems like a solved problem. Many collocation-finders exist, so one might suspect that most could suffice for finding MWU dictionary headwords. To verify this, we evaluate nine existing collocation-finders to see which best identifies valid headwords. We evaluate using two completely separate gold standards: (1) WordNet and (2) a compendium of Internet dictionaries. Although web-based resources are dynamic and have better coverage than WordNet (especially for acronyms and names), we show that WordNet-based scores are comparable to those using Internet MRDs. Yet the evaluations indicate that significant improvement is still needed in MWU-induction.

As an attempt to improve MWU headword induction, we introduce several algorithms using Latent Semantic Analysis (LSA). LSA is a technique which automatically induces semantic relationships between words. We use LSA to try to eliminate proposed MWUs which are semantically compositional. Unfortunately, this does not help.

Yet when we use LSA to identify substitutable MWUs, we do show modest performance gains.

2 Previous Approaches

For decades, researchers have explored various techniques for identifying interesting collocations. There have essentially been three separate kinds of approaches for accomplishing this task. These approaches could be broadly classified into (1) segmentation-based, (2) word-based and knowledge-driven, or (3) word-based and probabilistic. We will illustrate strategies that have been attempted in each of the approaches. Since we assume knowledge of whitespace, and since many of the first and all of the second categories rely upon human input, we will be most interested in the third category.

2.1 Segmentation-driven Strategies

Some researchers view MWU-finding as a natural by-product of segmentation. One can regard text as a stream of symbols and segmentation as a means of placing delimiters in that stream so as to separate logical groupings of symbols from one another. A segmentation process may find that a symbol stream should not be delimited even though subcomponents of the stream have been seen elsewhere. In such cases, these larger units may be MWUs.

The principal work on segmentation has focused either on identifying words in phonetic streams (Saffran, et. al, 1996; Brent, 1996; de Marcken, 1996) or on tokenizing Asian and Indian languages that do not normally include word delimiters in their orthography (Sproat, et al, 1996; Ponte and Croft 1996; Shimohata, 1997; Teahan, et al., 2000; and many others). Such efforts have employed various strategies for segmentation, including the use of hidden Markov models, minimum description length, dictionary-based approaches, probabilistic automata, transformation-based learning, and text compression. Some of these approaches require significant sources of human knowledge, though others, especially those that follow data compression or HMM schemes, do not.

These approaches could be applied to languages where word delimiters exist (such as in European languages delimited by the space character). However, in such languages, it seems more prudent to simply take advantage of delimiters rather than introducing potential errors by trying to find word boundaries while ignoring knowledge of the delimiters. This suggests that in a language with whitespace, one might prefer to begin at the word level and identify appropriate word combinations.

2.2 Word-based, knowledge-driven Strategies

Some researchers start with words and propose MWU induction methods that make use of parts of speech, lexicons, syntax or other linguistic structure (Justeson and Katz, 1995; Jacquemin, et al., 1997; Daille, 1996). For example, Justeson and Katz indicated that the patterns NOUN NOUN and ADJ NOUN are very typical of MWUs. Daille also suggests that in French, technical MWUs follow patterns such as "NOUN de NOUN" (1996, p. 50). To find word combinations that satisfy such patterns in both of these situations necessitates the use of a lexicon equipped with part of speech tags. Since we are interested in knowledge-free induction of MWUs, these approaches are less directly related to our work. Furthermore, we are not really interested in identifying constructs such as general noun phrases as the above rules might generate, but rather, in finding only those collocations that one would typically need to define.

2.3 Word-based, Probabilistic Approaches

The third category assumes at most whitespace and punctuation knowledge and attempts to infer MWUs using word combination probabilities. Table 1 (see next page) shows nine commonly-used probabilistic MWU-induction approaches. In the table, f_X and P_X signify frequency and probability of a word X. A variable XY indicates a word bigram and ξ_{XY} indicates its expected frequency at random. An overbar signifies a variable's complement. For more details, one can consult the original sources as well as Ferreira and Pereira (1999) and Manning and Schütze (1999).

3 Lexical Access

Prior to applying the algorithms, we lemmatize using a weakly-informed tokenizer that knows only that whitespace and punctuation separate words. Punctuation can either be discarded or treated as words. Since we are equally interested in finding units like "Dr." and "U. S.," we opt to treat punctuation as words.

Once we tokenize, we use Church's (1995) suffix array approach to identify word *n*-grams that occur at least *T* times (for *T*=10). We then rank-order the

Table 1: Probabilistic Approaches

METHOD	FORMULA
Frequency (Guiliano, 1964)	f_{XY}
Pointwise Mutual Information (MI) (Fano, 1961; Church and Hanks, 1990)	$\log_2 (P_{XY} / P_X P_Y)$
Selectional Association (Resnik, 1996)	$\frac{P_{X\|Y} * MI_{XY}}{\sum_Z Pr_{Z\|Y} * MI_{ZY}}$
Symmetric Conditional Probability (Ferreira and Pereira, 1999)	$P_{XY}{}^2 / P_X P_Y$
Dice Formula (Dice, 1945)	$2 f_{XY} / (f_X + f_Y)$
Log-likelihood (Dunning, 1993; Daille, 1996)	$-2\log\frac{[P_X P_Y P_{\overline{X}} P_{\overline{Y}}]^{f_Y}}{[P_{XY} P_{\overline{XY}}]^{f_{XY}} [P_{X\overline{Y}} P_{\overline{XY}}]^{f_{\overline{X}Y}}}$
Chi-squared (χ^2) (Church and Gale, 1991)	$\sum_{\substack{i\in\{X,\overline{X}\} \\ j\in\{Y,\overline{Y}\}}} \frac{(f_{ij} - \xi_{ij})^2}{\xi_{ij}}$
Z-Score (Smadja, 1993; Fontenelle, et al., 1994)	$\frac{f_{XY} - \xi_{XY}}{\sqrt{\xi_{XY}(1-(\xi_{XY}/N))}}$
Student's t-Score (Church and Hanks, 1990)	$\frac{f_{XY} - \xi_{XY}}{\sqrt{f_{XY}(1-(f_{XY}/N))}}$

n-gram list in accordance to each probabilistic algorithm. This task is non-trivial since most algorithms were originally suited for finding two-word collocations. We must therefore decide how to expand the algorithms to identify general *n*-grams (say, $C=w_1 w_2 ... w_n$). We can either generalize or approximate. Since generalizing requires exponential compute time and memory for several of the algorithms, approximation is an attractive alternative.

One approximation redefines X and Y to be, respectively, the word sequences $w_1 w_2 ... w_i$ and $w_{i+1} w_{i+2} ... w_n$, where *i* is chosen to maximize $P_X P_Y$. This has a natural interpretation of being the expected probability of concatenating the two most probable substrings in order to form the larger unit. Since it can be computed rapidly with low memory costs, we use this approximation.

Two additional issues need addressing before evaluation. The first regards document sourcing. If an *n*-gram appears in multiple sources (eg., Congressional Record versus Associated Press), its likelihood of accuracy should increase. This is particularly true if we are looking for MWU headwords for a general versus specialized dictionary. Phrases that appear in one source may in fact be general MWUs, but frequently, they are text-specific units. Hence, precision gained by excluding single-source *n*-grams may be worth losses in recall. We will measure this trade-off.

Second, evaluating with punctuation as words and applying no filtering mechanism may unfairly bias against some algorithms. Pre- or post-processing of *n*-grams with a linguistic filter has shown to improve some induction algorithms' performance (Daille, 1996). Since we need knowledge-poor induction, we cannot use human-suggested filtering rules as in Section 2.2. Yet we can filter by pruning *n*-grams whose beginning or ending word is among the top *N* most frequent words. This unfortunately eliminates acronyms like "U. S." and phrasal verbs like "throw up." However, discarding some words may be worthwhile if the final list of *n*-grams is richer in terms of MRD headwords. We therefore evaluate with such an automatic filter, arbitrarily (and without optimization) choosing *N*=75.

4 Evaluating Performance

A natural scoring standard is to select a language and evaluate against headwords from existing dictionaries in that language. Others have used similar standards (Daille, 1996), but to our knowledge, none to the extent described here. We evaluate thousands of hypothesized units from an unconstrained corpus. Furthermore, we use two separate evaluation gold standards: (1) WordNet (Miller, et al, 1990) and (2) a collection of Internet MRDs. Using two gold standards helps valid MWUs. It also provides evaluation using both static and dynamic resources. We choose to evaluate in English due to the wealth of linguistic resources.

Table 2: Outputs from each algorithm at different sorted ranks

Rank	ZScore	$\chi 2$	SCP	Dice	Mutual Info.	Select Assoc.	Log Like.	TScore	Freq
1	Iwo Jima	Buenos Aires	Buenos Aires	Buenos Aires	Iwo Jima	United States	United States	United States	United States
2	bona fide	Iwo Jima	Iwo Jima	Iwo Jima	bona fide	House of Repre-sentatives	Los Angeles	Los Angeles	Los Angeles
4	Burkina Faso	Suu Kyi	Suu Kyi	Suu Kyi	Wounded Knee	Los Angeles	New York	New York	New York
8	Satanic Verses	Sault Ste	Sault Ste	Sault Ste	Hubble Space Telescope	my colleagues	Soviet Union	my colleagues	my colleagues
16	Ku Klux	Ku Klux	Ku Klux	Ku Klux	alma mater	H . R	Social Security	High School	High School
32	Pledge of Allegiance	Pledge of Allegiance	Pledge of Allegiance	Pledge of Allegiance	Coca - Cola	War II	House of Repre-sentatives Wednesday	* *	* *
64	Telephone & amp ; Telegraph	Telephone & amp ; Telegraph	Telephone & amp ; Telegraph	Internal Revenue	Planned Parent-hood	Prime Minister	* * *	real estate	New Jersey
128	Prime Minister	Prime Minister	Prime Minister	Salman Rushdie	Sault Ste . Marie	both sides	At the same time	Wall Street	term care
256	Lehman Hutton	Lehman Hutton	Lehman Hutton	tongue - in - cheek	o ’ clock	At the same	del Mar	all over	grand jury
512	La Habra	La Habra	La Habra	compens-atory and punitive	20th - Century	Monday night	days later	80 percent	Great Northern
1024	telephone interview	telephone interview	telephone interview	Food and Agriculture	Sheriff ’s deputies	South Dakota	County Jail	where you	300 million

The “* *” and “* * *” are actual units.

In particular, we use a randomly-selected corpus consisting of a 6.7 million word subset of the TREC databases (DARPA, 1993-1997).

Table 2 illustrates a sample of rank-ordered output from each of the different algorithms (following the cross-source, filtered paradigm described in section 3). Note that algorithms in the first four columns produce results that are similar to each other as do those in the last four columns. Although the mutual information results seem to be almost in a class of their own, they actually are similar overall to the first four sets of results; therefore, we will refer to the first five columns as “information-like.” Similarly, since the last four columns share properties of the frequency approach, we will refer to them as “frequency-like.”

One’s application may dictate which set of algorithms to use. Our gold standard selection reflects our interest in general word dictionaries, so results we obtain may differ from results we might have obtained using terminology lexicons.

If our gold standard contains K MWUs with corpus frequencies satisfying threshold (T=10), our figure of merit (FOM) is given by

$$\frac{1}{K}\sum_{i=1}^{K} P_i,$$

where P_i (precision at i) equals i/H_i , and H_i is the number of hypothesized MWUs required to find the i^{th} correct MWU. This FOM corresponds to area under a precision-recall curve.

4.1 WordNet-based Evaluation

WordNet has definite advantages as an evaluation resource. It has in excess of 50,000 MWUs, is freely accessible, widely used, and is in electronic form. Yet, it obviously cannot contain every MWU. For instance, our corpus contains 177,331 *n*-grams (for $2 \le n \le 10$) satisfying $T \ge 10$, but WordNet contains only 2610 of these. It is unclear, therefore, if algorithms are wrong when they propose MWUs that are not in WordNet. We will assume they are wrong but with a special caveat for proper nouns. WordNet includes few proper noun MWUs. Yet several algorithms produce large numbers of proper nouns. This biases against them. One could contend that all proper nouns MWUs are valid, but we disagree. Although such may be MWUs, they are not necessarily MRD headwords; one would not include every proper noun in a dictionary, but rather, those needing definitions. To overcome this, we will have two scoring modes. The first, "S" mode (standing for *some*) discards any proposed capitalized *n*-gram whose uncapitalized version is not in WordNet. The second mode "N" (for *none*) disregards all capitalized *n*-grams.

Table 3 illustrates algorithmic performance as compared to the 2610 MWUs from WordNet. The first double column illustrates "out-of-the-box" performance on all 177,331 possible *n*-grams. The second double column shows cross-sourcing: only hypothesizing MWUs that appear in at least two separate datasets (124,952 in all), but being evaluated against all of the 2610 valid units. Double columns 3 and 4 show effects from high-frequency filtering the *n*-grams of the first and second columns (reporting only 29,716 and 17,720 *n*-grams) respectively.

As Table 3 suggests, for every condition, the information-like algorithms seem to perform best at identifying valid, general MWU headwords. Moreover, they are enhanced when cross-sourcing is considered; but since much of their strength comes from identifying proper nouns, filtering has little or even negative impact. On the other hand, the frequency-like approaches are independent of data source. They also improve significantly with filtering. Overall, though, after the algorithms are judged, even the best score of 0.265 is far short of the maximum possible, namely 1.0.

Table 3: WordNet-based scores

Prob algo-rithm	(1) WordNet		(2) WordNet cross-source		(3) WordNet +Filter		(4) WordNet cross-source +Filter	
	S	N	S	N	S	N	S	N
Zscore	.222	.146	.263	.193	.220	.129	.265	.173
SCP	.221	.145	.262	.192	.220	.129	.265	.173
Chi-sqr	.222	.146	.263	.193	.220	.129	.265	.173
Dice	.242	.167	.265	.199	.230	.142	.256	.172
MI	.191	.122	.245	.169	.185	.111	.233	.151
SA	.057	.051	.058	.053	.182	.125	.202	.143
Loglike	.049	.050	.068	.064	.118	.095	.177	.129
T-score	.050	.051	.050	.052	.150	.109	.160	.118
Freq	.035	.037	.034	.037	.144	.105	.152	.112

4.2 Web-based Evaluation

Since WordNet is static and cannot report on all of a corpus' *n*-grams, one may expect different performance by using a more all-encompassing, dynamic resource. The Internet houses dynamic resources which *can* judge practically every induced *n*-gram. With permission and sufficient time, one can repeatedly query websites that host large collections of MRDs and evaluate each *n*-gram.

Having approval, we queried: (1) onelook.com, (2) acronymfinder.com, and (3) infoplease.com. The first website interfaces with over 600 electronic dictionaries. The second is devoted to identifying proper acronyms. The third focuses on world facts such as historical figures and organization names.

To minimize disruption to websites by reducing the total number of queries needed for evaluation, we use an evaluation approach from the information retrieval community (Sparck-Jones and van Rijsbergen, 1975). Each algorithm reports its top 5000 MWU choices and the union of these choices (45192 possible *n*-grams) is looked up on the Internet. Valid MWUs identified at *any* website are assumed to be the *only* valid units in the data.

Algorithms are then evaluated based on this collection. Although this strategy for evaluation is not flawless, it is reasonable and makes dynamic evaluation tractable. Table 4 shows the algorithms' performance (including proper nouns).

Though Internet dictionaries and WordNet are completely separate "gold standards," results are surprisingly consistent. One can conclude that WordNet may safely be used as a gold standard in future MWU headword evaluations. Also,

Table 4: Performance on Internet data

Prob algorithm	(1) Internet	(2) Internet cross-source	(3) Internet +Filter	(4) Internet cross-source +Filter
Z-Score	.165	.260	.169	.269
SCP	.166	.259	.170	.270
Chi-sqr	.166	.260	.170	.270
Dice	.183	.258	.187	.267
MI	.139	.234	.140	.234
SA	.027	.033	.107	.194
Log Like	.023	.043	.087	.162
T-score	.025	.027	.110	.142
Freq	.016	.017	.104	.134

one can see that Z-scores, χ^2, and SCP have virtually identical results and seem to best identify MWU headwords (particularly if proper nouns are desired). Yet there is still significant room for improvement.

5 Improvement strategies

Can performance be improved? Numerous strategies could be explored. An idea we discuss here tries using induced semantics to rescore the output of the best algorithm (filtered, cross-sourced Zscore) and eliminate semantically compositional or modifiable MWU hypotheses.

Deerwester, et al (1990) introduced Latent Semantic Analysis (*LSA*) as a computational technique for inducing semantic relationships between words and documents. It forms high-dimensional vectors using word counts and uses singular value decomposition to project those vectors into an optimal *k*-dimensional, "semantic" subspace (see Landauer, et al, 1998).

Following an approach from Schütze (1993), we showed how one could compute latent semantic vectors for any word in a corpus (Schone and Jurafsky, 2000). Using the same approach, we compute semantic vectors for every proposed word *n*-gram $C=X_1X_2...X_n$. Since LSA involves word counts, we can also compute semantic vectors (denoted by $\boldsymbol{\Omega}$) for *C*'s subcomponents. These can either include ($\{X_i\}_{i=1}^{n}$) or exclude ($\{X_i^*\}_{i=1}^{n}$) *C*'s counts. We seek to see if induced semantics can help eliminate incorrectly-chosen MWUs. As will be shown, the effort using semantics in this nature has a very small payoff for the expended cost.

5.1 Non-compositionality

Non-compositionality is a key component of valid MWUs, so we may desire to emphasize *n*-grams that are semantically non-compositional. Suppose we wanted to determine if *C* (defined above) were non-compositional. Then given some meaning function, Ψ, *C* should satisfy an equation like:

$$g(\ \Psi(C)\ ,\ h(\ \Psi(X_1),...,\Psi(X_n)\ )\ ) \geq 0, \qquad (1)$$

where *h* combines the semantics of *C*'s subcomponents and *g* measures semantic differences. If *C* were a bigram, then if *g*(a,b) is defined to be |a-b|, if h(c,d) is the sum of c and d, and if $\Psi(e)$ is set to -log P_e, then equation (1) would become the pointwise mutual information of the bigram. If g(a,b) were defined to be $(a-b)/b^{1/2}$, and if h(a,b)=ab/N and $\Psi(X)=f_X$, we essentially get Z-scores. These formulations suggest that several of the probabilistic algorithms we have seen include non-compositionality measures already. However, since the probabilistic algorithms rely only on distributional information obtained by considering juxtaposed words, they tend to incorporate a significant amount of non-semantic information such as syntax. Can semantic-only rescoring help?

To find out, we must select *g*, *h*, and Ψ. Since we want to eliminate MWUs that are compositional, we want *h*'s output to correlate well with *C* when there is compositionality and correlate poorly otherwise. Frequently, LSA vectors are correlated using the cosine between them:

$$\cos(\boldsymbol{X},\boldsymbol{Y}) = \frac{\boldsymbol{X}\cdot\boldsymbol{Y}}{\|\boldsymbol{X}\|\,\|\boldsymbol{Y}\|}\ .$$

A large cosine indicates strong correlation, so large values for g(**a**,**b**)=1-|cos(**a**,**b**)| should signal weak correlation or non-compositionality. *h* could

represent a weighted vector sum of the components' semantic vectors with weights (w_i) set to either 1.0 or the reciprocal of the words' frequencies.

Table 5 indicates several results using these settings. As the first four rows indicate and as desired, non-compositionality is more apparent for $\mathbf{\Omega}_X$* (i.e., the vectors derived from excluding C's counts) than for $\mathbf{\Omega}_X$. Yet, performance overall is horrible, particularly considering we are rescoring Z-score output whose score was 0.269. Rescoring caused five-fold degradation!

Table 5: Equation 1 settings

g(**a**,**b**)	h(**a**)	Ψ(X)	w_i	Score on Internet
1-|cos(**a**,**b**)|	$\sum_{i=1}^{n} w_i a_i$	$\mathbf{\Omega}_X$	1	0.0517
			1/fi	0.0473
		$\mathbf{\Omega}_X$*	1	0.0598
			1/fi*	0.0523
|cos(**a**,**b**)|		$\mathbf{\Omega}_X$	1	0.174
			1/fi	0.169
		$\mathbf{\Omega}_X$*	1	0.131
			1/fi*	0.128

What happens if we instead *emphasize* compositionality? Rows 5-8 illustrate the effect: there is a significant recovery in performance. The most reasonable explanation for this is that if MWUs and their components are strongly correlated, the components may rarely occur except in context with the MWU. It takes about 20 hours to compute the $\mathbf{\Omega}_X$* for each possible *n*-gram combination. Since the probabilistic algorithms already identify *n*-grams that share strong distributional properties with their components, it seems imprudent to exhaust resources on this LSA-based strategy for non-compositionality.

These findings warrant some discussion. Why did non-compositionality fail? Certainly there is the possibility that better choices for *g*, *h*, and Ψ could yield improvements. We actually spent months trying to find an optimal combination as well as a strategy for coupling LSA-based scores with the Z-scores, but without avail. Another possibility: although LSA can find semantic relationships, it may not make semantic decisions at the level required for this task. This seems to be a significant component. Yet there is still another: maybe semantic compositionality is not always bad. Interestingly, this is often the case. Consider *vice_president*, *organized crime*, and *Marine_Corps*. Although these are MWUs, one would still expect that the first is related to *president*, the second relates to *crime*, and the last relates to *Marine*. Similarly, tokens such as *Johns_Hopkins* and *Elvis* are anaphors for *Johns_Hopkins_University* and *Elvis_Presley*, so they should have similar meanings.

This begs the question: can induced semantics help at all? The answer is "yes." The key is using LSA where it does best: finding things that are similar — or substitutable.

5.2 Non-substitutivity

For every collocation $C=X_1X_2..X_{i-1}\mathbf{X_i}X_{i+1}..X_n$, we attempt to find other similar patterns in the data, $X_1X_2..X_{i-1}\mathbf{Y}X_{i+1}..X_n$. If X_i and Y are semantically related, chances are that *C* is substitutable.

Since LSA excels at finding semantic correlations, we can compare $\mathbf{\Omega}_{Xi}$ and $\mathbf{\Omega}_Y$ to see if *C* is substitutable. We use our earlier approach (Schone and Jurafsky, 2000) for performing the comparison; namely, for every word W, we compute $\cos(\mathbf{\Omega}_{w,}\mathbf{\Omega}_R)$ for 200 randomly chosen words, R. This allows for computation of a correlaton mean (μ_W) and standard deviation (σ_W) between W and other words. As before, we then compute a normalized cosine score ($\overline{\cos}$) between words of interest, defined by

$$\overline{\cos}(X_i,Y) = \min_{k\in\{X_i,Y\}} \frac{\cos(\mathbf{\Omega}_{Xi},\mathbf{\Omega}_Y)-\mu_k}{\sigma_k}.$$

With this set-up, we now look for substitutivity. Note that phrases may be substitutable and *still* be headword if their substitute phrases are themselves MWUs. For example, *dioxide* in *carbon_dioxide* is semantically similar to *monoxide* in *carbon_monoxide*. Moreover, there are other important instances of valid substitutivity:

- Abbreviations
  Al≡Albert ⇒ *Al_Gore≡Albert_Gore*
- Morphological similarities
  Rico≡Rican ⇒ *Puerto_Rico≡Puerto_Rican*
- Taxonomic relationships
  bachelor≈master⇒
  bachelor_'_s_degree≈master_'_s_degree.

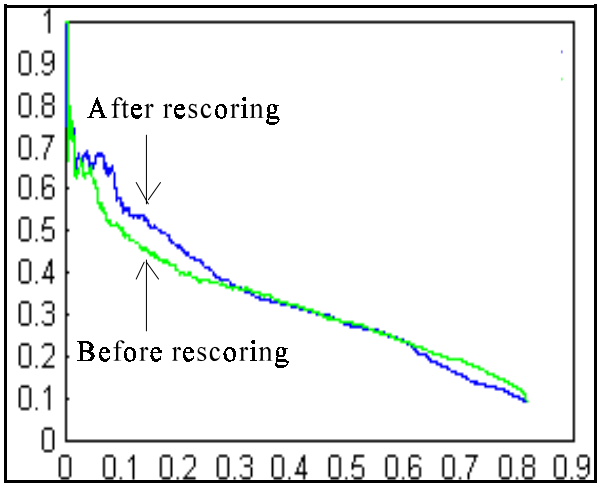
However, *guilty* and *innocent* are semantically related, but *pleaded_guilty* and *pleaded_innocent* are not MWUs. We would like to emphasize only *n*-grams whose substitutes are valid MWUs.

To show how we do this using LSA, suppose we want to rescore a list *L* whose entries are potential MWUs. For every entry X in L, we seek out all other entries whose sorted order is less than some maximum value (such as 5000) that have all but one word in common. For example, suppose X is "bachelor_'_s_degree." The only other entry that matches in all but one word is "master_'_s_degree." If the semantic vectors for "bachelor" and "master" have a normalized cosine score greater than a threshold of 2.0, we then say that the two MWUs are in each others substitution set. To rescore, we assign a new score to each entry in substitution set. Each element in the substitution set gets the same score. The score is derived using a combination of the previous Z-scores for each element in the substitution set. The combining function may be an averaging, or a computation of the median, the maximum, or something else. The maximum outperforms the average and the median on our data. By applying in to our data, we observe a small but visible improvement of 1.3% absolute to .282 (see Fig. 1). It is also possible that other improvements could be gained using other combining strategies.

6 Conclusions

This paper identifies several new results in the area of MWU-finding. We saw that MWU headword evaluations using WordNet provide similar results to those obtained from far more extensive web-based resources. Thus, one could safely use WordNet as a gold standard for future evaluations. We also noted that information-like algorithms, particularly Z-scores, SCP, and $\chi 2$, seem to perform best at finding MRD headwords regardless of filtering mechanism, but that improvements are still needed. We proposed two new LSA-based approaches which attempted to address issues of non-compositionality and non-substitutivity. Apparently, either current algorithms already capture much non-compositionality or LSA-based models of non-compositionality are of little help. LSA *does* help somewhat as a model of substitutivity. However, LSA-based gains are small compared to the effort required to obtain them.

Figure 1: Precision-recall curve for rescoring



Acknowledgments

The authors would like to thank the anonymous reviewers for their comments and insights.

References

AcronymFinder.com(2000-1).http://www.acronymfinder.com. Searches between March 2000 and April 2001.

Brent, M.R. and Cartwright, T.A. (1996). Distributional regularity and phonotactic constraints are useful for segmentation. *Cognition*, 61, 93-125.

Choueka, Y. (1988). Looking for needles in a haystack or locating interesting collocation expressions in large textual databases. *Proceedings of the RIAO*, pp. 38-43.

Church, K.W. (1995). *N-grams*. Tutorial at ACL, '95. MIT, Cambridge, MA.

Church, K.W., & Gale, W.A. (1991). Concordances for parallel text. *Proc. of the 7th Annual Conference of the UW Center for ITE New OED & Text Research*, pp. 40-62, Oxford.

Church, K.W., & Hanks, P. (1990). Word association norms, mutual information and lexicography. *Computational Linguistics*, Vol. 16, No. 1, pp. 22-29.

Daille, B. (1996). "Study and Implementation of Combined Techniques from Automatic Extraction of Terminology" Chap. 3 of "The Balancing Act": Combining Symbolic and Statistical Approaches to Language (Klavans, J., Resnik, P. (eds.)), pp. 49-66

DARPA (1993-1997). DARPA text collections: *A.P. Material, 1988-1990*, *Ziff Communications Corpus, 1989*, *Congressional Record of the 103rd Congress*, and *Los Angeles Times*.

Deerwester, S., S.T. Dumais, G.W. Furnas, T.K. Landauer, and R. Harshman. (1990) Indexing by Latent Semantic Analysis. *Journal of the American Society of Information Science*, Vol. 41

de Marcken, C. (1996) *Unsupervised Language*

Acquisition, Ph.D., MIT

Dias, G., S. Guilloré, J.G. Pereira Lopes (1999). Language independent automatic acquisition of rigid multiword units from unrestricted text corpora. *TALN*, Cargèse.

Dice, L.R. (1945). Measures of the amount of ecologic associations between species. *Journal of Ecology*, 26, 1945.

Dunning, T (1993). Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*. Vol. 19, No. 1.

Fano, R. (1961). *Transmission of Information*. MIT Press, Cambridge, MA.

Finke, M. and Weibel, A. (1997) Speaking mode dependent pronunciation modeling in large vocabulary conversational speech recognition. *Eurospeech-97.*

Ferreira da Silva, J., Pereira Lopes, G. (1999). A local maxima method and a fair dispersion normalization for extracting multi-word units from corpora. *Sixth Meeting on Mathematics of Language*, pp. 369-381.

Fontenelle, T., Brüls, W., Thomas, L., Vanallemeersch, T., Jansen, J. (1994). DECIDE, MLAP-Project 93-19, deliverable D-1a: Survey of collocation extraction tools. Tech. Report, Univ. of Liege, Liege, Belgium.

Giuliano, V. E. (1964) "The interpretation of word associations." In M.E. Stevens et al. (Eds.) Statistical association methods for mechanized documentation, pp. 25-32. National Bureau of Standards Miscellaneous Publication 269, Dec. 15, 1965.

Gregory, M. L., Raymond, W.D., Bell, A., Fosler-Lussier, E., Jurafsky, D. (1999). The effects of collocational strength and contextual predictability in lexical production. *CLS99*, University of Chicago.

Heid, U. (1994). On ways words work together. *Euralex-99.*

Hindle, D. (1990). Noun classification from predicate-argument structures. *Proceedings of the Annual Meeting of the ACL*, pp. 268-275.

InfoPlease.com (2000-1). http://www.infoplease.com. Searches between March 2000 and April 2001.

Jacquemin, C., Klavans, J.L., & Tzoukermann, E. (1997). Expansion of multi-word terms for indexing and retrieval using morphology and syntax. *Proc. of ACL* 1997, Madrid, pp. 24-31.

Justeson, J.S. and S.M.Katz (1995). Technical terminology: some linguistic properties and an algorithm for identification in text. *Natural Language Engineering* 1:9-27.

Kilgariff, A., & Rose, T. (1998). Metrics for corpus similarity & homogeneity. Manuscript, ITRI, University of Brighton.

Landauer, T.K., P.W. Foltz, and D. Laham. (1998) Introduction to Latent Semantic Analysis. *Discourse Processes*. Vol. 25, pp. 259-284.

Manning, C.D., Schütze, H. (1999) *Foundations of Statistical Natural Language Processing*, MIT Press, Cambridge, MA, 1999.

Mikheev, A., Finch, S. (1997). Collocation lattices and maximum entropy models. WVLC, Hong Kong.

Miller, G. (1990).“WordNet: An on-line lexical database,” International Journal of Lexicography, 3(4).

OneLook.com (2000-1). http://www.onelook.com. Searches between March 2000 and April 2001.

Ponte, J.M., Croft, B.W. (1996). Useg: A Retargetable word segmentation procedure for information retrieval. Symposium on Document Analysis and Information Retrieval ‘96. Technical Report TR96-2, University of Massachusetts.

Resnik, P. (1996). Selectional constraints: an information-theoretic model and its computational realization. *Cognition*. Vol. 61, pp. 127-159.

Saffran, J.R., Newport, E.L., and Aslin, R.N. (1996). Word segmentation: the role of distributional cues. *Journal of Memory and Language*, Vol. 25, pp. 606-621.

Schone, P. and D. Jurafsky. (2000) Knowledge-free induction of morphology using latent semantic analysis. *Proc. of the Computational Natural Language Learning Conference*, Lisbon, pp. 67-72.

Schütze, H. (1993) Distributed syntactic representations with an application to part-of-speech tagging. *Proceedings of the IEEE International Conference on Neural Networks*, pp. 1504-1509.

Shimohata, S., Sugio, T., Nagata, J. (1997). Retrieving collocations by co-occurrences and word order constraints. *Proceedings of the 35th Annual Mtg. of the Assoc. for Computational Linguistics*. Madrid. Morgan-Kauffman Publishers, San Francisco. Pp. 476-481.

Smadja, F. (1993). Retrieving collocations from text: Xtract. *Computational Linguistics*, 19:143-177.

Sparck-Jones, K., C. van Rijsbergen (1975) Report on the need for and provision of an “ideal” information retrieval text collection, British Library Research and Development Report, 5266, Computer Laboratory, University of Cambridge.

Sproat R, Shih, C. (1990) A statistical method for finding word boundaries in Chinese text. *Computer Processing of Chinese & Oriental Languages*, Vol. 4, No. 4.

Sproat, R. (1992) *Morphology and Computation*. MIT Press, Cambridge, MA.

Sproat, R.W., Shih, C., Gale, W., Chang, N. (1996) A stochastic finite-state word segmentation algorithm for Chinese. *Computational Linguistics*, Vol. 22, #3.

Teahan, W.J., Yingyin, W. McNab, R, Witten, I.H. (2000). A Compression-based algorithm for Chinese word segmentation. *ACL* Vol. 26, No. 3, pp. 375-394.