# A Powerful and General Approach to Context Exploitation in Natural Language Processing

**Robert W. Means[1*], Syrus C. Nemat-Nasser[1], Adrian T. Fan[1], and Robert Hecht-Nielsen[2,1]**

[1]Fair Isaac Corporation
3661 Valley Centre Drive
San Diego, CA 92130
[*]rwm@fairisaac.com

[2]Computational Neurobiology
Institute for Neural Computation
ECE Department
University of California, San Diego
La Jolla, CA 92093-0407
rh-n@ucsd.edu

## Abstract

In natural language, the meaning of a lexeme often varies due to the specific surrounding context. Computational approaches to natural language processing can benefit from a reliable, long-range-context-dependent representation of the meaning of each lexeme that appears in a given sentence. We have developed a general new technique that produces a context-dependent 'meaning' representation for a lexeme in a specific surrounding context. The 'meaning' of a lexeme in a specific context is represented by a list of *semantically replaceable elements* the members of which are other lexemes from our experimental lexicon. We have performed experiments with a lexicon composed of individual English words and also with a lexicon of individual words and selected phrases. The resulting lists can be used to compare the 'meaning' of conceptual units (individual words or frequently-occurring phrases) in different contexts and also can serve as features for machine learning approaches to classify semantic roles and relationships.

## 1 Introduction

Statistical natural language approaches build models based on annotated corpora as well as unlabeled corpora. The latter, requiring unsupervised knowledge acquisition, has the advantage of larger training sets—it is possible to exploit corpora composed of billions of words. A number of researchers have observed that such use of very large corpora improves the stability of statistical models (e.g. Banko and Brill, 2001).

The mathematical procedures employed here are based upon Hecht-Nielsen's neuroscience theory of cognition (Hecht-Nielsen, 2003). In a nutshell, this theory holds that cognition is based upon a procedure of ruling out all unreasonable conclusions and then deciding, of the remaining conclusions, which are the least worst ones. This mathematical symbolic predictive technique is called *confabulation*. The knowledge employed by confabulation is vast quantities of conditional probabilities for pairs of symbols. This knowledge, which is of no value for reasoning or probabilistic inference, is readily obtainable. Hecht-Nielsen's discovery is that, given the proper coding of a problem into symbols, confabulation works essentially as well as reasoning would if we were in possession of the necessary 'omniscient' knowledge that reasoning requires. Unfortunately, 'omniscient' knowledge is not practically obtainable, thereby making attempts to implement reasoning, in any form, futile. Confabulation, on the other hand, although it does require storage and use of large volumes of knowledge, is simple and practical (e.g., see Table 5 for the number of items of knowledge used in the experiments reported here). Confabulation provides an explicit mechanism that can now be used to build artificial intelligence.

Our approach to 'meaning' representation for lexemes is to provide a set of similar elements that are grammatically and/or semantically interchangeable with a given lexeme. Others have constructed lexical similarity clusters using order-dependent co-occurrence statistics, particularly with N-gram models—see Brown et al. (1992) for an example where words are sorted into exclusive classes based on bigram statistics. The occurrence statistics of bigrams do stabilize for frequent words given a training corpus of hundreds of millions of words. However, beyond tri-grams, the theoretical size of a training corpus required for completeness is unreasonable. Our method uses only pairwise conditionals.

To analyze a given text stream, we use a hierarchy consisting of a *word-level representation* and a *concep-*

*tual-unit-level representation* to analyze arbitrary single-clause English sentences. Each of these representations uses a lexicon of language element tokens to encode free text as described below. The representation of a sentence with two levels of hierarchy at the word level and the phrase level is consistent with Late Assignment of Syntax Theory, an analysis by synthesis model advocated by Townsend and Bever (2001).

## 2 Lexicon Construction

We construct a case-sensitive *word-level lexicon* based on frequency of occurrence in our large English text corpus of approximately 100 million sentences containing more than 2.3 billion white-space-separated tokens. The raw corpus was assembled from a number of newswire corpora, spanning roughly 14 years beginning in 1988, and hand-selected modern-English, after 1800, Gutenberg texts. We limit our lexicon to 63,000 tokens at which point the frequency rank corresponds to a minimum of 1000 occurrences.

After construction of our *word-level lexicon*, we construct a *postword word-level* knowledge base for use in creating a *conceptual-unit lexicon*. To create this word-level knowledge base, we count token bigram occurrences within our corpus and then calculate *antecedent support* conditional probabilities as follows: For a given token $t_i$ representing the $i^{th}$ word in our lexicon, for each word lexicon token $t_j$ that occurs immediately following $t_i$ in the training corpus, the antecedent support probability is approximated as:

$$p(t_i \mid t_j) \cong c(t_i, t_j) / c(t_j) \qquad (1)$$

where $c(t_i, t_j)$ is the count of the times the $j^{th}$ word follows the $i^{th}$ word in the corpus and $c(t_j)$ is the total count of the $j^{th}$ word in the corpus, excluding occurrences immediately following a punctuation mark. Based on these quantities, 'meaningful' knowledge is identified and assigned non-zero weights in the postword knowledge base if it has a co-occurrence count $c(t_i, t_j) \geq 3$ and antecedent support probability $p(t_i \mid t_j) > 1.0 \times 10^{-4}$. Approximately 17 million token-to-token knowledge items satisfied these two conditions.

We compose our *conceptual-unit lexicon* from the 63,000 *word-level* tokens plus an additional 63,000 automatically identified *conceptual units*, each consisting of between two and five word tokens. Conceptual units are identified using the pairwise *postword word-level* knowledge base as follows for each sentence in the training corpus:

- Assume the $i^{th}$ word of a sentence starts a conceptual unit;
- As long as p($i^{th}$ word| ($i^{th}$+1) word) > $T_0$, the conceptual unit continues up to a maximum length;
- Punctuation marks, such as commas and quotation marks terminate a conceptual unit directly.

The maximum conceptual unit length and the threshold $T_0$ have been somewhat arbitrarily chosen as 5 and 0.02 respectively. We implement a complete frequency sort of all observed conceptual units in the corpus. All conceptual units with a minimum of 1000 occurrences are retained. These 63,000 additional tokens are added to the word level lexicon resulting in a conceptual unit lexicon with 126,000 unique tokens. Figure 1 illustrates the segmentation of an example sentence into word-level tokens and conceptual-unit-level tokens.

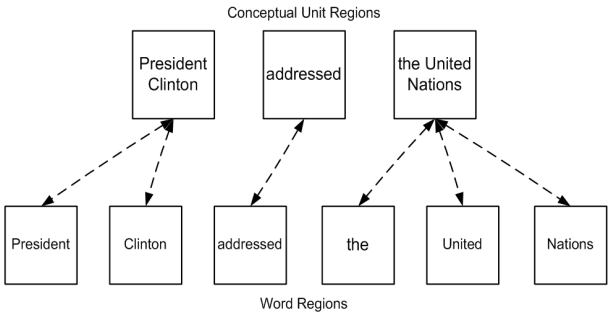


**Figure 1**. Segmentation of a sentence into word tokens and conceptual unit tokens

## 3 SRE Expansion

A Semantically Replaceable Element (SRE) is a word or conceptual unit that can be used as a grammatically-consistent, semantically similar substitute in a given linguistic context. An SRE is similar to a synonym. However, words and conceptual units are rarely exact synonyms and often have multiple meanings that only become clear in context. Our SRE expansion method uses knowledge derived from the entire training corpus to produce a list of 'synonyms' and then uses specific surrounding context in a sentence to prune this list of candidates into a list of SREs.

SRE expansion proceeds as follows: A test sentence without internal punctuation is presented to the system. This sentence is represented twice, once as a sequence of individual word tokens and once as a sequence of conceptual unit tokens (Figure 1). Figure 2 illustrates the hierarchical architecture used for SRE expansion. The hierarchy has two layers: a word analysis layer and a conceptual unit analysis layer. We create knowledge bases between the tokens in the conceptual unit layer and the tokens in the word layer in the same manner described for the *postword word-level* knowledge base.

A conceptual unit has connections both to and from its postwords and prewords. Separate knowledge bases to and from the conceptual unit layer are created for both postwords and prewords of conceptual units out to a distance of plus or minus two words (see Figure 2). These knowledge bases are normalized to limit the dynamic range of the strengths. Normalization proceeds as follows:

- If $t_i$ is not followed by $t_j$ at least 3 times in our corpus, the knowledge item is discarded;
- If $p(t_i \mid t_j)$ is less than or equal to a threshold $T_l = 1.0 \times 10^{-4}$, the knowledge item is discarded;
- The strength $W_{ji}$ to token $t_j$ from token $t_i$ is calculated as $W_{ji} = \log_2(p(t_i \mid t_j)/T_l)$.

Logarithmic scaling of the antecedent support probability reflects a biologically-inspired compression of dynamic range.

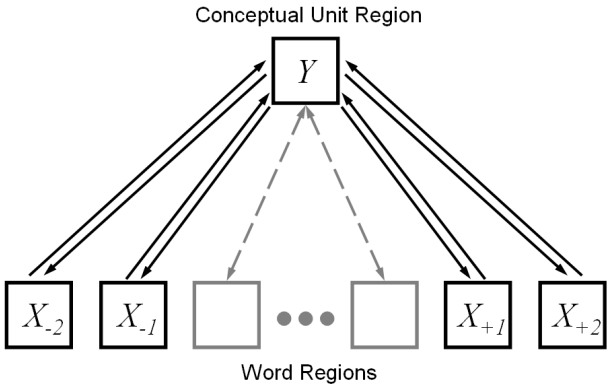


**Figure 2**. The hierarchical knowledge architecture: One conceptual unit representation region is used for SRE expansion along with two preceding word regions and two postword regions. Solid arrows indicate independent pairwise unidirectional knowledge bases. Dashed arrows indicate the correspondence between a conceptual unit and the individual word tokens from which it is composed.

The knowledge bases between the conceptual unit layer and the word layer are used to create a list of potential synonyms. This is done by activating a token for the $i^{th}$ conceptual unit in the sentence in the conceptual unit region (*Y* in Figure 2). The conceptual-unit-to-word knowledge bases activate other tokens in the four preword and postword regions ($X_{-2}$, $X_{-1}$, $X_{+1}$, and $X_{+2}$ in Figure 2). Each token within these regions is activated with the strength $W_{ji}$. Those word tokens, in turn, activate tokens back in the conceptual unit region by means of the word-to-conceptual-unit knowledge bases. The result is a set of active tokens in the original conceptual unit region that are potential synonyms. This process does not rely on the specific sentence context; it uses the knowledge bases, trained on the entire corpus, to produce candidate synonyms. For example, when a word (e.g. "suit") is placed on the conceptual unit region, its preword and postword tokens are 'excited' in the word regions below with strength of excitation equal to the corresponding weights. Those words in turn excite potential synonyms that have most potential senses in the conceptual unit region (e.g. lawsuit, jacket). The first fourteen potential synonyms are listed in Table 1. Other senses of "suit" are also excited with strengths that depend on their usage in the training corpus.

| **suit** |
|---|
| suits |
| lawsuit |
| jacket |
| shirt |
| pants |
| lawsuits |
| jackets |
| trousers |
| coat |
| shirts |
| sweater |
| blazer |
| slacks |
| civil suit |

**Table 1**. The first fourteen potential synonyms of the conceptual unit "suit"

To perform SRE expansion for a given sentence, we first generate a list of up to 100 candidate synonyms for each conceptual unit—*It is possible though rare for a word token to return less than 100 potential synonyms using the procedure described above.* The words surrounding the conceptual unit are then used to remove entries, pruning the list of potential synonyms. We use up to two prewords and two postwords. Due to edge effects at the start and end of the sentence, we always have 2, 3, or 4 context words. The pruning operation proceeds in two steps: First, we count the number of knowledge base connections from the surrounding context words to the actual word in the sentence; these items of knowledge must be present in the word-to-conceptual unit knowledge bases (Figure 2). Second, we 'confirm' potential synonyms that receive an equal or greater number of connections from the surrounding context words. The pruned list is termed an SRE expansion. It tends to have semantic and syntactic agreement with the given conceptual unit.

| **Apple** | **filed** | **a suit** | **against** | **IBM** |
|---|---|---|---|---|
| Sun Microsystems | had filed | a lawsuit | against Microsoft | AT&T |
| Compaq | alleges | a civil suit | versus | Intel |
| Intel | dismissed | a complaint | was filed | Intel Corp. |
| IGM | settled | the suit | vs. | HewlettPackard |
| Sun | to drop | lawsuits | filed | Dell |
| Microsoft | copyright | suits | alleging | Microsoft |
| Lotus | | the lawsuit | accusing | Oracle |
| Digital | | suits | that gave | Motorola |
| Microsoft Corp. | | classaction lawsuit | struggle against | Sony |
| Intel Corp. | | a petition | in federal court | Apple Computer |
| Computer | | an appeal | were filed | General Motors |
| Power | | a motion | charging | General Electric |
| AST | | a claim | against Yugoslavia's | NEC |
| Genentech | | civil suits | that ended | Digital |
| International Business Machines | | lawsuit | was sparked | 3M |
| Ascend | | in a suit | that followed | American Express |
| MCI | | a class action | brought | Philip Morris |
| AT&T | | in a lawsuit | to oust | Procter & Gamble |
| Motorola | | the complaint | stemming from | Kodak |

**Table 2**. SRE expansion example: the word “suit” as in lawsuit. The first nineteen expansion terms are displayed.

| **He wore** | **a suit** | **to the** | **wedding** |
|---|---|---|---|
| Wearing | the suit | to his | birthday |
| wearing | suits | to their | bridal |
| wore | a jacket | to our | funeral |
| wears | a coat | to the traditional | graduation |
| who wore | a white | to his own | marriage |
| was wearing | a shirt | to the military | gala |
| and wearing | a black | to her | cocktail |
| who wears | a gray | to a | Wedding |
| donned | a helmet | to my | Christmas |
| to wear | a T-shirt | to your | mourning |
| wear | camouflage | | lavish |
| don | | | inaugural |
| is wearing | | | black-tie |
| donning | | | festive |
| his trademark | | | coronation |
| he wore | | | prom |
| shirt | | | glittering |
| jacket | | | chiffon |
| trademark | | | evening |

**Table 3**. SRE expansion example: the word “suit” as in clothing. The first nineteen expansion terms are displayed.

| These | arbitrarily | chosen | phrases | demonstrate | our | meaning | representation |
|---|---|---|---|---|---|---|---|
| Those | unfairly | chose | words | demonstrated | one's | significance | representations |
| Many | randomly | constructed | language | to demonstrate | to our | truth | protections |
| The two | automatically | shaped | songs | demonstrates | my | purpose | protection |
| A few | strictly | | themes | demonstrating | on our | motives | treatment |
| They | deliberately | | symbols | illustrate | people's | interpretation | distribution |
| You | properly | | rhetoric | indicate | our commitment | dimension | expression |
| The first | they have been | | sentences | have demonstrated | their | sense | approximation |
| | carefully | | images | have shown | the government's | motives | images |
| | should not be | | poems | prove | your | nature | participation |
| | who have been | | words | confirm | its commitment | dimensions | and democratic |
| | should be | | remarks | suggest | America's | expression | description |
| | are being | | the word | reveal | of their | phrase | supervision |
| | correctly | | names | underscore | to their | truths | recognition |
| | appropriately | | to describe | show | the ability | insight | status |
| | were being | | texts | to prove | the administration's | identity | constituency |
| | they will be | | scenes | assess | the president's | emotion | voting |
| | they had been | | colors | underline | their skills | themes | equations |
| | routinely | | comments | reflect | Washington's | message | immunity |
| | selectively | | | doubts about | the party's | vitality | disclosure |

**Table 4**. SRE expansion example: an arbitrary sentence.

| Knowledge Base | Items of Knowledge |
|---|---|
| $Y$ to $X_{-2}$ | 16,432,495 |
| $Y$ to $X_{-1}$ | 16,189,554 |
| $Y$ to $X_{+1}$ | 13,594,106 |
| $Y$ to $X_{+2}$ | 16,796,927 |
| $X_{-2}$ to $Y$ | 22,451,444 |
| $X_{-1}$ to $Y$ | 22,089,368 |
| $X_{+1}$ to $Y$ | 17,597,506 |
| $X_{+2}$ to $Y$ | 23,973,514 |

**Table 5**. Size of knowledge bases used for the SRE expansion

The SRE expansion procedure was applied to 33 sentences which contained a total of 233 words. Each word had 100 possible synonyms. The average number of confirmed synonyms due to the surrounding context was 28.2 with a standard deviation of 35.7. Tables 2, 3, and 4 present three example sentences that have been expanded using our method—a maximum of nineteen expansion terms are displayed.

## 4 Discussion

Our SRE expansion method provides a context-specific 'meaning' representation providing application builders with features that could be applied to problems including word sense disambiguation and named entity recognition. Miller et al. (2004) describe a relevant technique for the latter. To quantify the quality of our SRE expansions will require an end-user application demonstration that we are unable to provide at this time.

Our approach uses a very large training corpus, a hierarchical architecture, and nine independent pairwise co-occurrence knowledge bases. Individually, these components have, in some form, been applied to computational natural language processing by other researchers. However, the combination of these components in our biologically-inspired framework has already produced novel methods that may prove useful to the computational linguistics community.

Our knowledge bases are large, but they are not exhaustive. Our confirmation method accommodates a certain amount of missing knowledge—instances where two language elements should be linked, but our training procedure has failed to identify this link. This approach is a compromise reflecting the fact that our knowledge bases still need improvement. To fix deficiencies in our current knowledge bases, we require further development. We do not believe that a pure unsupervised statistical learning approach will suffice. Instead, we are working to develop 'education' procedures that apply supervised learning and hybrid learning techniques to improve the quality and completeness of our pairwise knowledge bases.

The authors wish to acknowledge significant past and present contributions to this project by Rion L. Snow and Katherine Mark.

## References


Banko, Michele and Brill, Eric, "Learning Curves and Natural Language Disambiguation", *Proceedings of HLT*, pp. 93-99, 2001.

Brown, P.F., V.J. Della Pietra, P.V. deSouza, J.C. Lai, and R.L. Mercer, "Class-based n-gram models of natural language", *Association for Computational Linguistics*, 1992.

Hecht-Nielsen, R., "A theory of thalamocortex" In: Hecht-Nielsen, R. and T. McKenna (Eds.), **Computational Models for Neuroscience**, pp. 85–124, London: Springer-Verlag, 2003.

Miller, S., Guineness, J., and A. Zamanian, "Name tagging with word clusters and discriminative training", To appear in *Proceedings of HLT*, 2004.

Townsend, David J. and Thomas G. Bever, **Sentence Comprehension**, The MIT Press, Cambridge MA, 2001.