Closed Yesterday and Closed Minds: Asking the Right Questions of the Corpus To Distinguish Thematic from Sentential Relations

Uri Zernik
Artificial Intelligence Laboratory
General Electric - Research and Development Center

Abstract

Collocation-based tagging and bracketing programs have attained promising results. Yet, they have not arrived at the stage where they could be used as pre-processors for full-fledged parsing. Accuracy is still not high enough.

To improve accuracy, it is necessary to investigate the points where statistical data is being misinterpreted, leading to incorrect results.

In this paper we investigate inaccuracy which is injected when a pre-processor relies solely on collocations and blurs the distinction between two separate relations: *thematic relations* and *sentential relations*.

Thematic relations are word pairs, not necessarily adjacent, (e.g., *adjourn* a *meeting*) that encode information at the concept level. Sentential relations, on the other hand, concern adjacent word pairs that form a noun group. E.g., *preferred stock* is a noun group that must be identified as such at the syntactic level.

Blurring the difference between these two phenomena contributes to errors in tagging of pairs such as *expressed concerns*, a verb-noun construct, as opposed to *preferred stocks*, an adjective-noun construct. Although both relations are manifested in the corpus as high mutual-information collocations, they possess different properties and they need to be separated.

In our method, we distinguish between these two cases by asking additional questions of the corpus. By definition, thematic relations take on further variations in the corpus. *Expressed concerns* (a thematic relation) takes *concerns expressed*, *expressing concerns*, *express his concerns* etc. On the other hand, *preferred stock* (a sentential relation) does not take any such syntactic variations.

We show how this method impacts pre-processing and parsing, and we provide empirical results based on the analysis of an 80-million word corpus. [1] [2]

Pre-Processing: The Greater Picture

Sentences in a typical newspaper story include idioms, ellipses, and ungrammatic constructs. Since authentic language defies textbook grammar, we must rethink our basic pars-

[1]This research was sponsored (in part) by the Defense Advanced Research Project Agency (DOD) and other government agencies. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Defense Advanced Research Project Agency or the US Government.

[2]We thank ACL/DCI (Data Collection Initiative), the Collins publishing company, and the Wall Street Journal, for providing invaluable on-line data.

[Separately/av] *comma*/cc [Kaneb/nm Services/nn] [said/vb] [holders/nn] [of/pp its/dt Class/nn A/aj **preferred/aj stock/nn**] *comma*/cc [failed/vb] [to/pp elect/vb] [two/aj directors/nn] [to/pp the/dt company/nn board/nn] when/cc [the/dt annual/aj meeting/nn] [resumed/vb] [Tuesday/aj] because/cc there/cc are/ax [questions/nn] as/cc [to/pp the/dt validity/nn] [of/pp the/dt proxies/nn] [submitted/vb] [for/pp review/nn] [by/pp the/dt group/nn] *period*/cc

[The/dt company/nn] [adjourned/vb] [its/pn annual/aj meeting/nn] May/nm 12/aj] [to/pp allow/vb] [time/nn] [for/pp negotiations/nn] and/cc **[expressed/vb] [concern/nn]** [about/pp future/aj actions/nn] [by/pp **preferred/vb holders/nn**] *period*/cc

Figure 1: Pre-processed Text Produced by NLcp

ing paradigm and tune it to the nature of the text under analysis.

Hypothetically, parsing could be performed by one huge unification mechanism [Kay, 1985; Shieber, 1986; Tomita, 1986] which would process sentences at any level of complexity. Such a mechanism would recieve its tokens in the form of words, characters, or morphemes, negotiate all given constraints, and produce a full chart with all possible interpretations.

However, when tested on a real corpus, (i.e., Wall Street Journal (WSJ) news stories), this mechanism fares poorly. For one thing, a typical well-behaved 34-word sentence produces hundreds of candidate interpretations. In effect the parsing burden is passed onto a post processor whose task is to select the appropriate parse tree within the entire forest.

For another, ill-behaved sentences – roughly one out of three WSJ sentences is problematic – yield no consistent interpretation whatsoever due to parsing failures.

To alleviate problems associated with rough edges in real text, a new strategy has emerged, involving text pre-processing. A pre-processor, capitalizing on statistical data [Church *et al.*, 1989; Zernik and Jacobs, 1990; Dagan *et al.*, 1991], and customized to the corpus itself, could abstract idiosyncracies, highlight regularities, and, in general, feed digested text into the unification parser.

What is Pre-Processing Up Against?

The Linguistic Phenomenon

Consider (Figure 1) a WSJ (August 19, 1987) paragraph processed by NLpc (NL corpus processing) [Zernik *et al.*, 1991]. Two types of linguistic constructs must be resolved by the preprocessor:

> Class A **preferred/AJ stock/NN** *comma*
>
> and **expressed/VB concern/NN** about

How can a program determine that preferred stock is an adjective-noun, while expressed concern is a verb-noun construct?

The Input

The scope of the pre-processing task is best illustrated by the input to the pre-processor shown in Figure 2.

This lexical analysis of the sentence is based on the Collins on-line dictionary (about 49,000 lexical entries extracted by NLpc) plus morphology. Each word is associated with *candidates* part of speech, and almost all words are ambiguous. The tagger's task is to resolve the ambiguity.

For example, ambiguous words such as services, preferred, and expressed, should be tagged as noun (*nn*), adjective (*aj*), and verb (*vb*), respectively. While some pairs (e.g., *annual meeting*) can be resolved easily, other pairs

Separately	AV	Kaneb	NM	Services	NN VB
said	AJ VB	holders	NN	of	PP
its	DT	Class	AJ NN	A	DT AJ
preferred	AJ VB	stock	NN VB	failed	AD VB
to	PP	elect	VB	two	AJ NN
directors	NN	to	PP	the	DT
company	NN	board	NN VB	when	CC
annual	AJ	meeting	NN VB	resumed	AJ VB
tuesday	NM	questions	NN VB	validity	NN
proxies	NN	submitted	AJ VB	group	NN VB

Figure 2: Lexical Analysis of Sentence: Words plus Part of Speech

(e.g., *preferred stock* and *expressed concerns*) are more difficult, and require statistical training.

Part-Of-Speech Resolution

The program can bring to bear 3 types of clues:

Local context: Consider the following 2 cases where local context dominates:

1. **the** preferred stock raised
2. **he** expressed concern about

The words *the* and *he* dictate that preferred and expressed are adjective and verb respectively. This kind of inference, due to its local nature, is captured and propagated by the pre-processor.

Global context: Global-sentence constraints are shown by the following two examples:

1. and preferred stock sold yesterday **was** ...
2. and expressed concern about ... ***period***

In case 1, a main verb is found (i.e., *was*), and *preferred* is taken as an adjective; in case 2, a main verb is not found, and therefore *expressed* itself is taken as the main verb. This kind of ambiguity requires full-fledged unification, and it is not handled by the pre-processor. Fortunately, only a small percent of the cases (in newspaper stories) depend on global reading.

Corpus-based preference: Corpus analysis (WSJ, 80-million words) provides word-association preference [Beckwith *et al.*, 1991]

collocation	total	vb-nn	aj-nn
preferred stock	2314	100	0
expressed concern	318	1	99

The construct *expressed concern*, which appears 318 times in the corpus, is 99% a verb-noun construct; on the other hand, *preferred stock*, which appears in the corpus 2314 times, is 99% an adjective-noun construct.[3]

Where Is The Evidence?

The last item, however, is not directly available. Since the corpus is not a-priori tagged, there is no direct evidence regarding part-of-speech. All we get from the corpus are numbers that indicate the mutual information score (MIS) [Church *et al.*, 1991] of collocations (9.9 and 8.7, for preferred stock and expressed concern, respectively). It becomes necessary to infer the nature of the combination from indirect corpus-based statistics as shown by the rest of this paper.

[3] For expository purposes we chose here two extreme, clear-cut cases; other pairs (e.g., *promised money*) are not totally biased towards one side or another.

Inferring Syntax from Collocations

In this section we describe the method used for eliciting word-association preference from the corpus.

Initial Observation: Co-occurrence Entails Sentential Relations

The basic intuition used invariably by all existing statistical taggers is stated as follows: Significant collocations (i.e., high MIS) predict syntactic word association. Since, for example, *preferred stock* is a significant collocation (mis 9.9), with all other clues assumed neutral, it will be marked as an integral noun group in the sentence.

However, is high mis always a good predictor? Figure 3 provides mutual information scores for *preferred, expressed,* and *closed* right collocations.

The first column (*preferred*) suggests mis is a perfect predictor. A count in the corpus confirms that a predictor based on collocations is always correct. A small sample of *preferred* collocations in context is given Figure 4. Notice that in all cases, *preferred* is an adjective.

Next Observation: Co-occurrence Entails Thematic Relations

While column 1 (*preferred*) yields good syntactic associations, column 2 (*expressed*) and column 3 (*closed*) yield different conclusions. It turns out (see Figure 4) that *expressed* collocations, even collocations with high mis, produce a bias towards false-positive groupings.[4]

If these collocation do not signify word groupings, what do they signify? An observation of *expressed* right collocates reveals that the words *surprise, confidence, skepticism, optimism, disappointment, support, hope, doubt, worry, satisfaction*, etc., are all thematic relations of *express.*

Namely, a pair such as *expressed disappointment* denotes an action-object relation which could come in many variants. The last part of Figure 4 shows various combinations of *express* and its collocates.

[4]Word associations based on corpus do not *dictate* the nature of word groupings; they merely provide a predictor that is accounted for with other local-context clues.

Using Additional Evidence

In light of this observation, it is necessary to test in the corpus whether collocations are fixed or variable. For a collocation word1-word2, if word1 and word2 combine in multiple ways, then word1-word2 is taken as a thematic relation; otherwise it is taken as a fixed noun group.

This test for *express*-word is shown in Figure 5. Each row provides the number of times each variant is found. Variants for *expressed concerns*, for example, are *concern expressed, express concern, expresses concern,* and *expressing concern*. Not shown here is the count for split co-occurrence [Smadja, 1991], i.e., express its concern, concern was expressed. The last column sums up the result as a ratio (variability ratio) against the original collocation.

In conclusion, for 12 out of 15 of the checked collocations we found a reasonable degree of variability.

Making Statistics Operational

While the analysis in Figure 5 provides the motivation for using additional evidence, we have two steps to take to make this evidence useful within an operational tagger.

Dealing with Small Numbers

Although the table in Figure 5 is adequate for expository purposes, in practice the different collected figures are spread over too many rubrics, making the numbers susceptible to noise.

To avoid this problem we short-cut the calculation above and collect all the co-occurrence of

9.9	preferred	stock	11.9	expressed	disappointment	20.4	closed	friday
9.8	preferred	dividend	11.6	expressed	skepticism	17.4	closed	monday
8.1	preferred	share	10.8	expressed	optimism	16.3	closed	tuesday
7.4	preferred	method	10.8	expressed	reservations	16.0	closed	thursday
7.4	preferred	holders	10.1	expressed	doubt	16.0	closed	today
7.0	preferred	stockholders	10.0	expressed	surprise	15.7	closed	wednesday
7.0	preferred	shareholders	10.0	expressed	satisfaction	15.5	closed	saturday
6.1	preferred	issue	9.6	expressed	confidence	13.8	closed	tomorrow
5.2	preferred	units	8.9	expressed	shock	13.8	closed	mouthed
5.0	preferred	series	8.8	expressed	hope	8.1	closed	minded
4.7	preferred	equity	8.7	expressed	concern	8.0	closed	caption
4.6	preferred	closed	8.7	expressed	worry	7.7	closed	milieu
4.5	preferred	customer	8.6	expressed	relief	7.5	closed	doors
4.1	preferred	course	8.2	expressed	interest	7.4	closed	yesterday
3.7	preferred	product	7.0	expressed	support	6.8	closed	dumps

Figure 3: Right-Collocations for Preferred, Expressed, and Closed

the roots of the words under analysis. Instead of asking: "what are the individual variants?" we ask "what is the total co-occurrence of the root pair?". For *expressed concerns* we check the incidence of *express-interest* (and of *interest-express*).

As a result, we get the lump sum without summing up the individual numbers.

Incorporating Statistics in Tagging

Co-occurence information regarding each pair of words is integrated, as described in Section 2.3, with other local-context clues. Thus, the fact that statistics provide a strong preference can always be overidden by other factors.

they preferred stock ...
the expressed interest by shareholders was ...

In both these cases the final call is dictated by syntactic markers in spite of strong statistical preference.

Conclusions

NLpc processes collocations by their category. In this paper, we investigated specifically the PastParticiple–Noun category (e.g., preferred-stock, expressed-concerns, etc.). Other categories (in particular ContinuousVerb–Noun as in *driving cars* vs. *operating systems*) are processed in a similar way, using slightly different evidence and thresholds.

The Figures

Total cases:	2031
Applicable cases:	400
Insufficient data:	23
Incorrect tagging:	19
Correct tagging:	358

Evaluation

Out of 2031 tagging cases counted, the algorithm was called in 400 cases. 1631 cases were not called since they did not involve collocations (or involved trivial collocations such as *expressed some fears.*) Out of 400 collocations the program avoided ruling in 23 cases due to insufficient data. Within the 377 tagged cases, 358 (94.9%) cases were correct, and 19 were incorrect.

90% Accuracy is Not Enough

Existing pre-processors [Church *et al.*, 1989; Zernik *et al.*, 1991] which have used corpus-based collocations, have attained levels of ac-

GE for the 585,000 shares of its	preferred	stock outstanding *period* The e
ume payments of dividends on the	preferred	stock in January *period* It sus
ohawk but lowered ratings on its	preferred	stock and commercial paper *comm
n* 3 from BAA *hyphen* 2 *comma*	preferred	stock to ba *hyphen* 2 from BAA
llar* 26.65 a share *period* The	preferred	is convertible until 5 P.M.. EDT
ares of common for each share of	preferred	*r-paren* *period* Cash will be
0 *pc* of Varity *ap* common and	preferred	shares outstanding *period* The
ng of up to *dollar* 250 million	preferred	shares *period* Terms of the tra
erms of the transaction call for	preferred	holders *comma* who previously a
sal *comma* to swap one share of	preferred	stock for 1.2 shares of common s
i *dollar* 2 million annually in	preferred	dividends *period* Artra owns 68
p* notes and 7,459 Lori series C	preferred	shares with a carrying value of
a share of newly issued series A	preferred	stock with a value equal to *dol
ance an adjustable *hyphen* rate	preferred	stock whose auction failed recen

id he told the house Mr. Dingell	expressed	concern *comma* sources said *co
ggested that the U.S. Mr. Harper	expressed	confidence that he and Mr. Baum
ne tax *period* Some legislators	expressed	concern that a gas *hyphen* tax
soybeans and feed grains *comma*	expressed	outrage over the case *comma* sa
bid *dash* *dash* *dash* GE unit	expressed	interest in financing offer for
hallenge *period* Mr. Wright has	expressed	dismay that a foreign company co
bt about their bank one also had	expressed	interest in Mcorp *ap* mvestment
italy *ap* President Cossiga and	expressed	concern about an Italian firm su
comma saying warner executives	expressed	surprise at Sony *ap* move but d
secretary Robert Mosbacher have	expressed	concern about the EC *ap* use of
thor on the nature paper *comma*	expressed	disappointment that he was not i
eber who *comma* he said *comma*	expressed	support for the idea *period* Ca

ving gold in the street and then	expressing	surprise when thieves walk by t
said that National Pizza Co. has	expressed	renewed interest in acquiring th
r. nixon *comma* Chinese leaders	expressed	no regret for the killings *comm
e Bay Area *ap* pastry community	express	disbelief that Ms. Shere kept on
presidents also are expected to	express	support for the Andean nations w
its predecessor *period* It also	expressed	its commitment to a free *hyphe
related Services Co. people who	express	interest in the certificates rec
c chairman Seidman *comma* while	expressing	concerns *comma* also said the
* on a tour of asia *comma* also	expressed	a desire to visit China *period
ponsored the senate plan *comma*	expressed	some confidence that his plan w
the nine supreme court justices	expressed	varying degrees of dissatisfact
nd primerica in his eagerness to	express	his linguistic doubts to America
st few weeks alone *dash* *dash*	expressing	their relief after crossing in
iterally flipped his wig *comma*	expressing	delight at having an excuse to
that the newspaper company said	expresses	confidence in the outcome of a
who no longer feel they have to	express	their zeal on the streets *comma
icans writing to the hostages to	express	their grief and support *period*
en summoned to chairman Gonzalez	expresses	sympathy for Sen. Riegle *comma
riod* Frequently *comma* clients	express	interest in paintings but do not

Figure 4: PREFERRED, EXPRESSED, and (root) EXPRESS collocations in context

curacy as high as 90%. A simple calculation reveals that a 34-word sentence might contain some 1-2 errors on the average.

This error rate is too high. Since the pre-processor's job is to eliminate from consideration possible parse trees, if the appropriate parse is eliminated by the pre-processor at the outset, it will never be recovered by the parser. As shown in this paper, it is now necessary to investigate in depth how various linguistic phenomena are reflected by statistical data.

X	e'sed X		X e'sed	e's X	e'ses X	e'sing X	v. ratio		
	mis	no	no	no	no	no	n1	n2	r
disappointment	11.9	89	2	1	5	6	14	89	.16
skepticism	11.6	57		1		2	3	57	.05
optimism	10.8	49		3	1	4	8	49	.16
reservations	10.8	33		3	2	1	6	33	.18
doubt	10.1	63	2	1	5	4	13	63	.20
surprise	10.0	69	1	5	2	1	9	69	.13
satisfaction	10.0	14	1	2			3	14	.21
confidence	9.6	67		1	4	1	6	67	.09
shock	8.9	12		3		1	4	12	.33
hope	8.8	46		2	1	4	7	46	.15
concern	8.7	318	30	31	9	25	95	318	.30
worry	8.7	13	1	6	3	2	12	13	.92
relief	8.6	23					0	23	.00
interest	8.2	294	4	6	9	11	30	294	.10
support	7.0	46	1	5		3	9	46	.20

Figure 5: 5 Variant Collocations for Express

References

R. Beckwith, C. Fellbaum, D. Gross, and G. Miller. Wordnet: A lexical database organized on psycholinguistic principles. In U. Zernik, editor, *Lexical Acquisition: Exploiting On-Line Dictionary to Build a Lexicon*. Lawrence Erlbaum Assoc., Hissdale, NJ, 1991.

K. Church, W. Gale, P. Hanks, and D. Hindle. Parsing, word associations, and predicate-argument relations. In *Proceedings of the International Workshop on Parsing Technologies*, Carnegie Mellon University, 1989.

K. Church, W. Gale, P. Hanks, and D. Hindle. Using statistics in lexical analysis. In U. Zernik, editor, *Lexical Acquisition: Using On-Line Resources to Build a Lexicon*. Lawrence Erlbaum Associates, Hillsdale, NJ, 1991.

I. Dagan, A. Itai, and U. Schwall. Two languages are more informative than one. In *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics*, Berkeley, CA, 1991.

M. Kay. Parsing in Functional Unification Grammar. In D. Dowty, L. Kartunnen, and A. Zwicky, editors, *Natural Language Parsing: Psychological, Computational, and Theoretical Perspectives*. Cambridge University Press, Cambridge, England, 1985.

S. Shieber. *An Introduction to Unification-based Approaches to Grammar*. Center for the Study of Language and Information, Palo Alto, California, 1986.

F. Smadja. Macrocoding the lexicon with co-occurrence knowledge. In U. Zernik, editor, *Lexical Acquisition: Using On-Line Resources to Build a Lexicon*. Lawrence Erlbaum Associates, Hillsdale, NJ, 1991.

M. Tomita. *Efficient Parsing for Natural Language*. Kluwer Academic Publishers, Hingham, Massachusetts, 1986.

U. Zernik and P. Jacobs. Tagging for learning. In *COLING 1990*, Helsinki, Finland, 1990.

U. Zernik, A. Dietsch, and M. Charbonneau. Imtoolset programmer's manual. Ge-crd technical report, Artificial Intelligence Laboratory, Schenectady, NY, 1991.