Quantifying lexical influence: Giving direction to context

V Kripâsundar
kripa@cs.buffalo.edu

CEDAR & Dept. of Computer Science
SUNY at Buffalo
Buffalo NY 14260, USA

Abstract

The relevance of context in disambiguating natural language input has been widely acknowledged in the literature. However, most attempts at formalising the intuitive notion of context tend to treat the word and its context symmetrically. We demonstrate here that traditional measures such as mutual information score are likely to overlook a significant fraction of all co-occurrence phenomena in natural language. We also propose metrics for measuring directed lexical influence and compare performances.

Keywords: contextual post-processing, defining context, lexical influence, directionality of context

1 Introduction

It is widely accepted that context plays a significant role in shaping all aspects of language. Indeed, comprehension would be utterly impossible without the extensive application of contextual information. Evidence from psycholinguistic and cognitive psychological studies also demonstrates that contextual information affects the activation levels of lexical candidates during the process of perception (Weinreich, 1980; McClelland, 1987). Garvin (1972) describes the role of context as follows:

> [The meaning of] a particular text [is] not the system-derived meaning as a whole, but that part of it which is included in the contextually and situationally derived meaning proper to the text in question. (p. 69–70)

In effect, this means that the context of a word serves to restrict its sense.

The problem addressed in this research is that of improving the performance of a natural-language recogniser (such as a recognition system for handwritten or spoken language). The recogniser output typically consists of an ordered set of candidate words (word-choices) for each word position in the input stream. Since natural language abounds in contextual information, it is reasonable to utilise this in improving the performance of the recogniser (by disambiguating among the word-choices).

The word-choices (together with their confidence values) constitute a confusion set. The recogniser may further associate a confidence-value with each of its word choices to communicate finer resolution in its output. The language module must update these confidence values to reflect contextual knowledge.

2 Linguistic post-processing

The language module can, in principle, perform several types of "post-processing" on the word-candidate lists that the recogniser outputs for the different word-positions. The most promising possibilities are:

- re-ranking the confusion set (and assigning new confidence-values to its entries), and,
- deleting low-confidence entries from the confusion set (*after* applying contextual knowledge)

Several researchers in NLP have acknowledged the relevance of context in disambiguating natural language input ((Evett *et al.*, 1991); (Zernik, 1991); (Hindle & Rooth, 1993); (Rosenfeld, 1994)). In fact, the recent revival of interest in statistical language processing is partly because of its (comparative) success in modelling context. However, a theoretically sound definition of context is needed to ensure that such re-ranking and deleting of word-choices helps and not hinders (Gale & Church, 1990).

Researchers in information theory have come up with many inter-related formalisations of the ideas of context and contextual influence, such as mutual information and joint entropy. However, to our knowledge, all attempts at arriving at a theoretical basis for formalising the intuitive notion of context have treated the word and its context symmetrically.

Many researchers ((Smadja, 1991); (Śrihari & Baltus, 1993)) have suggested that the information-theoretic notion of mutual information score (MIS) directly captures the idea of context. However, MIS is deficient in its ability to detect one-sided correlations (*cf.* Table 1), and our research indicates that asymmetric influence measures are required to properly handle them (Kripàsundar, 1994).

For example, it seems quite unlikely that *any* symmetric information measure can accurately capture the co-occurrence relationship between the two words 'Paleolithic' and 'age' in the phrase 'Paleolithic age'. The suggestion that 'age' exerts as much influence on 'Paleolithic' as *vice versa* seems ridiculous, to say the least. What is needed here is a directed (*ie, one-sided*) influence measure (DIM), something that serves as a measure of influence of one word on another, rather than as a simple, symmetric, "co-existence probability" of two words. Table 1 illustrates how a DIM can be effective in detecting lexical and lexico-semantic associations.

3 Comparing measures of lexical influence

We used a section of the Wall Street Journal (WSJ) corpus containing 102K sentences (over two million words) as the training corpus for the partial results described here. The lexicon used was a simple 30K-word superset of the vocabulary of the training corpus.

The results shown here serve to strengthen our hypothesis that non-standard information measures are needed for the proper utilisation of linguistic context. Table 1 shows some pairs of words that exhibit differing degrees of influence on each other. It also demonstrates very effectively that one-sided information measures are much better than symmetric measures at utilising context properly. The arrow between each pair of words in the table indicates the direction of influence (or flow of information). The preponderance of word-pairs that exhibit only one direction of significant influence (*eg*, 'according'→'to') shows that no symmetric score could have captured the correlations in all of these phrases.

Our formulation of directed influence is still evolving. The word-pairs in Table 1 have been selected randomly from the test-set with the criterion that they scored "significantly" (*ie*, > 0.9) on at least one of the three measures D1, D2 and D3. The four measures (including MIS) are defined as follows:

$$\begin{aligned}
\mathrm{MIS}(w_1 w_2) &= \log\left(\frac{P(w_1 w_2)}{P(w_1)P(w_2)}\right) \\
\mathrm{D1}(w_1/w_2) &= \frac{P(w_1 w_2)}{P(w_2)} = \frac{\#w_1 w_2}{\#w_2} \\
\mathrm{D2}(w_1/w_2) &= \texttt{step1}\left(\frac{\#w_1 w_2}{\#\mathrm{Cmax}}\right) \times \mathrm{D1} \\
\mathrm{D3}(w_1/w_2) &= \texttt{step2}\left(\frac{\#w_1 w_2}{\#\mathrm{Cmax}}\right) \times \mathrm{D1}
\end{aligned}$$

In these definitions, $\#w_1 w_2$ denotes the frequency of co-occurrence of the words w_1 and w_2,[1] while $\#w_1$, and $\#w_2$ represent (respectively) the frequencies of their (unconditional) occurrence.

$\#\mathrm{Cmax} \overset{def}{=} \max\limits_{w_1 w_2}(\#w_1 w_2)$ is defined to be the maximum co-occurrence frequency in the corpus, and appears to be a better normalisation factor than the size of the corpus itself.

The definition of MIS implicitly incorporates the size of the corpus, since it has two $P()$ terms in the denominator, and only one in the numerator. The DIM's, on the other hand, have balanced fractions. Therefore, we have not included a log-term in the definitions of D1, D2, and D3 above.

D1 is a straightforward estimation of the conditional probability of co-occurrence. It forms a baseline for performance evaluations, but is prone to sparse data problems (Dunning, 1993).

The `step()` functions in D2 and D3 represent two attempts at minimising such errors. These functions are piecewise-linear mappings of the normalised co-occurrence frequency, and are used as scaling factors. Their effect is apparent in Table 1, especially in the bottom third of the table, where the low frequency of the primer pushes D3 down to insignificant levels.

The metrics D2 and D3 can and should be normalised, perhaps to the 0–1 range, in order to facilitate integration with other metrics such as the recogniser's confidence value. Similarly, the lack of normalisation of MIS hampers direct comparison of scores with the three DIM's.

4 Discussion

Of the several different types of word-level associations, lexical and lexico-semantic associations are among the most significant local associations. Lexical (or associative) context is characterised by rigid word order, and usually implies that the primer and the primed together act as one lexical unit. Lexico-semantic associations are exemplified by phrasal verbs (*eg*, 'fix up'), and are characterised by morphological complexity in the verb part and spatial flexibility in the phrase as a whole.

It is noteworthy that all the three DIM's capture the notions of lexical (*ie*, fixed) and lexico-semantic associations in one formula (albeit to differing degrees of success). Thus we have 'staff' and 'reporter' influencing each other almost equally, while the asymmetric influence on 'in' from its right context ('addition') is also detected by the DIM's.

It is our contention that symmetric measures constrain the re-ranking/proposing process significantly, since they are essentially blind to a significant fraction (perhaps *more than half*) of all co-occurrence phenomena in natural language.

5 Summary and Future Work

The preliminary results described in this work establish clearly that non-standard metrics of lexical

[1] Note that the exact word order of w_1 and w_2 is irrelevant here.

Word-pair w_L w_R			$(\#w_L, \#w_R, \#w_Lw_R)$	MIS	D1	D2	D3
new	←	york	(6927, 2697, 2338)	5.551	0.866	3.463	3.463
according	→	to	(1084, 54580, 1083)	3.629	0.999	2.996	2.996
staff	←	reporter	(1613, 1205, 1157)	7.111	0.960	2.879	2.879
staff	→	reporter	(1613, 1205, 1157)	7.111	0.717	2.150	2.150
new	→	york	(6927, 2697, 2338)	5.551	0.337	1.348	1.348
on	→	the	(13025, 116356, 3483)	1.554	0.267	1.334	1.334
vice	→	president	(1017, 2678, 784)	6.384	0.770	1.540	1.285
at	←	least	(11158, 795, 665)	5.039	0.836	1.671	1.247
compared	→	with	(585, 11362, 551)	5.139	0.941	1.881	1.244

Table 1: **Asymmetry in co-occurrence relationships:** Word-pairs with "significant" influence in either direction have been selected randomly from the test-set. Note that very few of these pairs exhibit comparable influence on each other. The arrows indicate the direction of lexical influence (or information flow). A DIM score of 1 or more implies a significant association, whereas an MIS below 4 is considered a chance association.

influence bear much promise. In fact, what we really need is a generalised information score, a measure that takes into account several factors, such as:

- directionality in correlation
- multiple words participating in a lexical relationship
- different (morphological) forms of words, and,
- spatial flexibility in the components of a collocation

The generalised information score would capture all the variations that are introduced by the above factors, and allow for the variants so as to reflect a "normalised" measure of contextual influence.

We have also been working with experimental measures which attach higher significance to the collocation frequency, (measures which, in essence, "trust" the recogniser more often). Our future work will involve bringing these various factors together into one integrated formalism.

References

Max Coltheart, editor. 1987. *Attention and Performance XII: The Psychology of Reading.* Lawrence Erlbaum.

Ted Dunning. 1993. Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, **19:1**:61–74.

LJ Evett, CJ Wells, FG Keenan, T Rose, and RJ Whitrow. 1991. Using linguistic information to aid handwriting recognition. *Proceedings of the International Workshop on Frontiers in Handwriting Recognition*, pages 303–311.

William A Gale and Kenneth W Church. 1990. Poor estimates of context are worse than none. In *Proceedings of the DARPA Speech and Natural Language Workshop*, pages 283–287.

Paul L Garvin. 1972. *On Machine Translation.* Mouton.

Donald Hindle and Mats Rooth. 1993. Structural ambiguity and lexical relations. *Computational Linguistics*, **19:1**:103–120.

V Kripȧsundar. 1994. *Drawing on Linguistic Context to Resolve Ambiguities* OR *How to imrove recongition in noisy domains.* Ph.D. thesis, Computer Science, SUNY@Buffalo. (*proposal*).

James L McClelland. 1987. The case for interactionism in language processing. In (Coltheart, 1987). Lawrence Erlbaum.

Ronald Rosenfeld. 1994. A hybrid approach to adaptive statistical language modeling. *Proceedings of the ARPA workshop on human language technology*, pages 76–81.

Frank Smadja. 1991. Macrocoding the lexicon with co-occurrence knowledge. *in* (Zernik, 1991), pages 165–190.

Rōhiṇi K Śrihari and Charlotte M Baltus. 1993. Use of language models in on-line recognition of handwritten sentences. *Proceedings of the Third International Workshop on Frontiers in Handwriting Recognition* (IWFHR III).

SN Śrihari, JJ Hull, and R Chaudhari. 1983. Integrating diverse knowledge sources in text recognition. *ACM Transactions on Office Information Systems*, **1:1**:68–87.

RM Warren. 1970. Perceptual restoration of missing speech sounds. *Science*, **167**:392–393.

Uriel Weinreich. 1980. *On Semantics.* University of Pennsylvania Press.

Uri Zernik, editor. 1991. *Lexical Acquisition: Exploiting On-line Resources to Build a Lexicon.* Lawrence Erlbaum.