

Automatic Extraction of High-Dimensional Semantics from Large Corpora: A Model of Human Syntactic Processing Constraints

Curt Burgess
Psychology Department
University of California
Riverside, CA 92521
curt@cassandra.ucr.edu
909-787-2392

Kevin Lund
Psychology Department
University of California
Riverside, CA 92521
kevin@cassandra.ucr.edu
909-787-2102

We present both a procedure whereby useful semantic information may be automatically extracted from corpora of conversational language and an experiment in which these semantic vectors support the empirical results of two parsing experiments.

We have developed a model of semantic representation, which we call HAL (Hyperspace Analogue to Language), that is derived from a 100-million word corpus. A ten-word window tracks co-occurrences of the most frequent 30,000 items. Within this ten-word window, co-occurrence values are inversely proportional to the number of words separating a specific pair. A word pair separated by a nine-word gap, for instance, would gain a co-occurrence strength of one, while the same pair appearing adjacent to one another would receive an increment of ten. Half of the 30,000 items are valid words, the other half are misspellings, numbers, etc. Tracking the co-occurrences produces a $30,000 \times 2$ matrix. Each row of the matrix may be extracted and is the co-occurrence vector for one word. To reduce the amount of data involved, the column variances of each vector is computed, and the 200 most variant columns are retained. These vectors can be viewed as the coordinates of points in high-dimensional space; each word occupies one point. This difference between two words' vector representations can be computed as the distance between the points defined by those vectors.

High-Dimensional Semantics and Human Syntactic Processing Constraints

Recent research has demonstrated that semantic constraints can affect syntactic processing of morphological verb ambiguities. Consider:

- 1a. The man paid by the parents was unreasonable.
- 1b. The ransom paid by the parents was unreasonable.

In (1a), *The man paid* has a bias for the simple past-tense. The disambiguating prepositional phrase, *by the parents*, results in slower reading times reflecting the initial misassignment of *The man* as an agent, rather than as patient. Sentence (1b) differs; the noun, *ransom*, can not be an agent. Burgess and Tanenhaus [BT] (1994) showed that sentences with a strong inanimate bias do not result in a slowing of reading time since the semantic information can be used by the parser. Ferreria and Clifton [FC] (1982) showed that semantic constraint made no difference in a similar study. BT

have demonstrated that the difference between the stimuli in these two studies was in part a difference in the constraint of the noun-verb pairs that were used.

We extracted 200-dimensional word vectors for all noun-verb pairs used by FC and BT. A Euclidean distance was calculated for each noun-verb pair using the logic that the closer the semantic distance between the noun and verb, the greater the semantic constraint. For example, *ransom paid* connotes a very specific situation, whereas, *man paid* is more general and less semantically constraining.

The semantic distance between the animate noun-verb pairs is greater than the distance between the inanimate noun-verb pairs for FC's stimuli. The same comparison for the BT stimuli shows no difference, reflecting the closer semantic relationship between the nouns and verbs. More critically, the BT noun-verb pairs have a shorter semantic distance than the FC stimuli. This shorter distance is consistent with the psycholinguistic results showing that the stimuli used in the BT experiment resulted in an elimination of the processing load associated with the syntactically-ambiguous reduced-relative construction. These results suggest that the semantic constraints manipulated in at least one class of psycholinguistic syntactic-processing experiments correspond to the word semantics that can be extracted from the high-dimensional model. Similar results from several other experiments will be presented at the conference.

Conclusions

We have presented results from HAL, a high-dimensional model of semantics that can be derived from large corpora. The procedural advantage of this methodology is that the text requires virtually no preprocessing and that the semantic vectors that are extracted seem to be very tolerant of the noise inherent in the conversational nature of the text. Nor does the method rely on human judgments about stimulus items. High-dimensional semantic vectors are an effective device for modeling the semantic constraints that can be used by the syntactic processor.

Acknowledgments

This work was supported by a NSF Presidential Faculty Fellow award and a UC Academic Senate grant to Curt Burgess.