Unsupervised Acquisition of Linguistic Knowledge

Dekang Lin

Department of Computer Science
University of Manitoba
Winnipeg, Manitoba, Canada, R3T 2N2
lindek@cs.umanitoba.ca

and

UMIACS
University of Maryland
College Park, Maryland, 20742
lindek@umiacs.umd.edu

While manually constructed treebanks have proven to be tremendously useful, they are seriously limited in size. We demonstrate a set of programs that learns from automatically parsed corpora in a completely unsupervised fashion. Although automatically parsed corpora contain much more noise than manually created treebanks, their virtually unlimited quantity may more than compensate for their lower quality.

This demonstration[1] consists of the following components:

- A broad-coverage parser called Minipar. Minipar is an efficient, principle-based parser was used to obtain the training data for the learning programs. It outputs the dependency structures of input sentences. Evaluation with the SUSANNE corpus show that about 88% of the dependency relationships obtained by Principar are correct.
- A collocation database. By extracting and aggregating the dependency relationships from the parsed corpus, we construct a collocation database. For any word that occurred in the corpus more than 10 times, the collocation database lists the frequency counts of other words that have direct dependency relation with the words in the corpus.
- Similarity matrices between about 12K nouns, 3.6K verbs and 5.6K adjective/adverbs. The similarity between a pair of words is obtained by treating the collocational context of the words as their features and use an information-theoretic measure to compute the similarity between two feature vectors. After the similarity matrices have been computed the top similar words of a given word are then clustered so that different clusters correspond to different senses of the word.
- Semantic frames of verbs. The parsed corpus, and the similarity matrices are then used to discover semantic frames of verbs. For example, the following are the automatically extracted semantic frames of the verb "appall". Each frame consists of a set of slots, such as subject (V:subj:N), object (V:comp1:N), and prepositional modifiers (V:at:N ...). A frame lists clusters of words that filled the slot in the parsed corpus.

```
(appall
  (V:subj:N action allegation bombing
    brutality campaign crackdown decision
    design discovery execution fighting
    greed idea image invasion killing move
    number practice prospect recommendation
    record remark report selection slaughter
    statement story style suggestion this
    threat use violence vote)
  (V:comp1:N American everyone executive
    family father he her I insider
    investor lawmaker leader life me
    member nation official others outsider
    people public resident scientist she
    We world)
  )
(appall
  (V:comp1:N American everyone executive
    family father he her I investor lawmaker
    leader life me member nation official
    others people public resident scientist
    she we world)
  (V:at:N abuse action change claim
    decision excess extent idea
    legislation notion prospect quality
    size statement strength thought
    timing)
  )
(appall
  (V:comp1:N American everyone father he
    her I me public she We)
  (V:at:N disregard hypocrisy)
  )
(appall
  (V:subj:N I she)
  (V:by:N action response)
  )
```

[1] Available at http://www.cs.umanitoba.ca/~lindek/nlldemo.htm