periment, all the words are first sorted based on its χ^2 with the keyword, and then the top 2,000 words are selected as trigger words.

ii) Parsing relationships associated with the keyword automatically decoded by a broad-coverage parser, with F-measure (i.e. the precision-recall combined score) at about 85% (reference temporarily omitted for the sake of blind review). The logical dependency relationships being utilized are listed below.

Noun: *subject-of,*
object-of,
complement-of,
has-adjective-modifier,
has-noun-modifier,
modifier-of,
possess,
possessed-by,
appositive-of

Verb: *has-subject,*
has-object,
has-complement,
has-adverb-modifier,
has-prepositional-phrase-modifier

Adjective: *modifier-of,*
has-adverb-modifier

Based on the above context features, the following three categories of context similarity features are defined:

(1) VSM-based (Vector Space Model based) trigger word similarity: the trigger words around the keyword are represented as a vector, and the word *i* in context *j* is weighted as follows:

$$weight(i, j) = tf(i, j) * \log \frac{D}{df(i)}$$

where $tf(i, j)$ is the frequency of word *i* in the *j-th* context; D is the number of documents in the pool; and $df(i)$ is the number of documents containing the word *i*. D and $df(i)$ are estimated using the document pool introduced above. The cosine of the angle between two resulting vectors is used as the context similarity measure.

(2) LSA-based (Latent Semantic Analysis based) trigger word similarity: LSA (Deerwester et al. 1990) is a technique used to uncover the underlying semantics based on co-occurrence data. The first step of LSA is to construct word-vs.-document co-occurrence matrix. Then singular value decomposition (SVD) is performed on this co-occurring matrix. The key idea of LSA is to reduce noise or insignificant association patterns by filtering the insignificant components uncovered by SVD. This is done by keeping only the top *k* singular values. By using the resulting word-vs.-document co-occurrence matrix after the filtering, each word can be represented as a vector in the semantic space.

In our experiment, we constructed the original word-vs.-document co-occurring matrix as follows: 100,000 documents from the TIPSTER corpus were used to construct the co-occurring matrix. We processed these documents using our POS tagger, and selected the top *n* most frequently mentioned words from each POS category as base words:

top 20,000 common nouns
top 40,000 proper names
top 10,000 verbs
top 10,000 adjectives
top 2,000 adverbs

In performing SVD, we set *k* (i.e. the number of nonzero singular values) as 200, following the practice reported in (Deerwester et al. 1990) and (Landauer & Dumais, 1997).

Using the LSA scheme described above, each word is represented as a vector in the semantic space. The co-occurring trigger words are represented as a vector summation. Then the cosine of the angle between the two resulting vector summations is computed, and used as the context similarity measure.

(3) LSA-based parsing relationship similarity: each relationship is in the form of $R_\alpha(w)$. Using LSA, each word *w* is represented as a