Help non-native English speakers improve their knowledge of colloquial English via idiom comprehension. Our system will help users learn idioms by 1) browsing definitions and examples to introduce them to new idioms and 2) returning idioms that match a given set of keywords the user searches for.

**SYSTEM DESIGN**

**Browsing and suggesting keywords:** Users can browse idioms (over 600 from the English Club dictionary and 100 most frequently used idioms) and their definitions alphabetically. Examples from the New York Times (over 1 million articles from 1987-2007) will be printed below. Users may suggest a keyword to associate with the idiom, allowing the system to learn more keywords for each idiom.

**Keyword matching:** Alternatively, users can input keywords to find idioms relating to the search term. Keywords are matched to the definition of the idiom, rather than the words in the idiom itself, allowing the user to search based on meanings. This is much more intuitive than existing systems which search strictly based on idioms, as an idiom’s meaning is many times unrelated to the terms which make up the idiom. Keywords are generated from user suggestions and two discovery algorithms.

**KEYWORD DISCOVERY ALGORITHMS**

Keywords were derived based on 1) definitions and 2) the context surrounding the idiom in the NYT examples through 2 different algorithms:

**Log likelihood**

\[ \text{Log likelihood : } \frac{P(t|T)}{P(t|NT)} \]

Terms for which \(-2\log(x) > 10\) are considered “topic signature terms” and were used as keywords.

**TF-idf**

\[ \text{Term frequency (tf): } \frac{1}{n} \text{ of the terms} \]

\[ \text{Inverse document frequency (idf): } \log(\text{no. of documents containing term}) \]

\[ \text{TF-idf : } \text{term frequency} \times \text{inverse document frequency} \]

Used five highest ranked words from each definition as the keywords.

**CONCLUSIONS**

Based on findings in Table 1, using the log likelihood method on the definitions produces the best keywords. However, because definitions are so short, keywords were not produced for all idioms. For those with matching idioms, the 2nd best algorithm was chosen, idf on the definitions.

Using the context of the idiom in the examples for finding keywords appeared to be inconclusive, providing non-sensical keywords.

**Final System**

We were able to create a system that is conducive to learning new idioms and browsing through a large set of idioms.