Chapter 3

Design

3.1 Algorithm

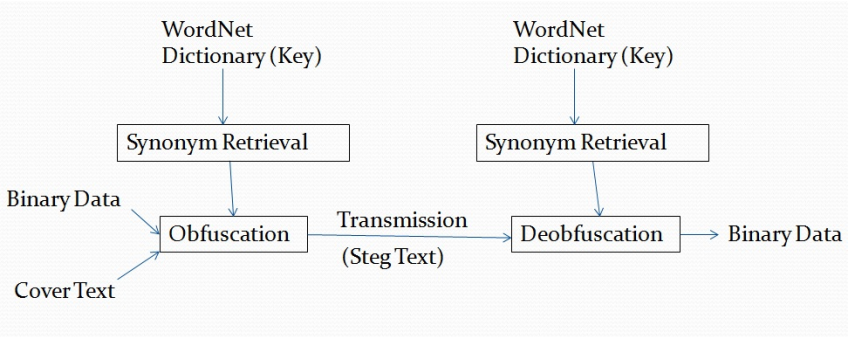

Figure 3.1: Figure showing structure of algorithm

3.1.1 Dictionary and Corpora

Dictionary

The dictionary that is used for the algorithm is the WordNet dictionary, produced by Princeton University in the US [20]. The dictionary contains around 150,000 words, organised into sets of synonyms, called synsets. A synset consists of the set of synonyms, a brief description (definition) of that sense, and in some cases one or two example sentences showing a typical usage of one of the words in the synset. There are around 115,000 synsets, some containing just a single word and some containing more than 8 synonyms. A search for a word in the dictionary database will return all synsets which contain that word, separated for the different word types (noun, verb, adverb and adjectives), and sorted by the frequency that they appear in some sample texts so the most common sense of a word is displayed first.