Lexical criminal identification for chatting corpus

ABSTRACT

This paper aims to identify lexical of criminal elements for chatting corpus, which involved suspect and victim conversation utterances. Lexical criminal identification requires three processes. The first is tokenization to automatically assign each lexical with a corresponding serial number in every suspect and victim utterance. The second is tagging the lexical with parts of speech to identify verbs and nouns in the utterances. The third is to identify and analyze the interrogative criminal construct to get the criminal evidence. The chatting corpus consists of 3,067 suspect and victim utterances with 16,278 words, collected from 9 criminal chatting cases. The results indicate that both verb and noun are the most important part of speech elements that represent the criminal constructs in chat utterances.