Automatic Knowledge Control System (AKCS)

Authors:
Naeim Rezaeian
Galina M. Novikova
Peoples Friendship University of Russia (RUDN)

Speacker’s
Naeim
Rezaeian
RUDN
Automatic Knowledge Control System

Knowledge control is a systematic process for collecting, analyzing, and interpreting information.

The learning process includes both planning and transfer of knowledge, as well as feedback support, which, first of all, consists in the control of knowledge. Carrier of knowledge, often considers the control process as an additional burden, taking away his time and energy from the educational process. This is an important issue today since the average number of students per teacher is growing, and the process of obtaining knowledge, as well as the process of evaluating the efficiency and quality of the acquired knowledge, have ceased to be limited in terms of time and space.

Automated knowledge control systems based on semantic textual analysis allow using various testing techniques that are aimed at determining the depth of understanding of the subject.
Methods of Knowledge control

1. Multiple choice:

Multiple choice questions are composed of one question with multiple possible answers (choices), including the correct answer and several incorrect answers.

2. Short answer:

Short answer questions are typically composed of a brief prompt that demands a written answer that varies in length from one or two words to a few sentences.

3. Essays:

Essay questions provide a complex prompt that requires written responses, which can vary in length from a couple of paragraphs to many pages.
Natural language processing (NLP) is a field of linguistics, computer science, and artificial intelligence in which computers analyze, understand, and derive meaning from human language in a smart and useful way.

- Information extraction
- Terminology extraction
- Summarizing
- Relation similarity
- Natural language generation
- Speech recognition
- Question answering
- Machine translation
- ....
Automatic Knowledge Control System (AKCS)

Architecture of AKCS

1- Data Input
- Current Answer
- Candidate Answer

2- Linguistic Pre-Processing
- Normalization
- Tokenization
- Lemmatization
- Stemming
- Formula Encoding

3- Distribution Semantics
- Word Embedded
- Similarity Module

Result
- Vector Similarity
- Weight of formula Encoding
Data Input

Questions were taken for the central control of knowledge in the discipline "Fundamentals of the development of corporate infocommunication systems".

The course is taught for students of such specialties as "Applied Informatics", "Information Technologies", "Business Informatics".

The data set as a whole contains 90 questions and 800 student responses. Responses were evaluated by experts using ratings from 0 to 100 percent.
Linguistic Pre-Processing

Normalization:
Removing punctuation marks, spaces between words, abbreviations, and etc.

Tokenization:
Given a character sequence and a defined document unit, tokenization is the task of chopping it up into pieces, called tokens, perhaps at the same time throwing away certain characters/words, such as punctuation.

Lemmatization:
Lemmatization is a more subtle process that uses vocabulary and morphological analysis and return the base or dictionary form of word, which is known as the lemma

Stemming:
A process of removing and replacing suffixes to get the root from of the word, which is called the stem
Formula Encoding

This part for replacing and eliminate the numbers and formula and foreign word in the input text. The algorithm replaces the elements of the program with the unit tokens. For example, every ID is replaced by the token < ID >. Or every numerical value with < value >. Now if a program has a statement in the form of $a = b + 10$, it will be replaced by the line or string of $< ID > = < ID > + < value >$.

Fingerprint Algorithm
Distribution Semantic

Character Based

- LCS
- Jaro
- Needleman Wunsch
- N-gram

Term Based

- Levenshtein
- Jaro-Winkler
- Smith Waterman
- Block Distance
- Dice’s Coefficient
- Jaccard Similarity
- Overlap Coefficient
- Cosine Similarity
- Euclidean Distance
- Matching Coefficient
CBO and Skip-gram

A two layer neural network to generate word embedding given a text corpus and return mapping of words in a vector space.

CBO Model:

\[ H_i = \{ i-h, ..., i-1, i+1, ..., i+h \} \]

\[
\begin{align*}
(w_i | w_{i-h}, ..., w_{i+h}) &= \frac{\exp(\varphi_{w_i} \sum_{j \in H_i} \theta_{w_j})}{\sum_w \exp(\varphi_w \sum_{j \in H_i} \theta_{w_j})} = \frac{1}{z_i} \prod_{j \in H_i} \exp \langle \varphi_{w_i}, \theta_{w_j} \rangle
\end{align*}
\]

Skip-gram Model:

\[
\begin{align*}
p(w_{i-h}, ..., w_{i+h} | w_i) &= \prod_{j \in H_i} p(w_j | w_i) = \prod_{j \in H_i} \frac{\exp \langle \varphi_{w_i}, \theta_{w_j} \rangle}{\sum_{w} \exp \langle \varphi_{w_j}, \theta_{w_i} \rangle} \\
&= \frac{1}{z_i} \prod_{j \in H_i} \exp \langle \varphi_{w_j}, \theta_{w_i} \rangle
\end{align*}
\]
Corpus

In linguistics, a corpus or text corpus is a large and structured set of texts. In corpus linguistics, they are used to do statistical analysis and hypothesis testing, checking occurrences or validating linguistic rules within a specific language territory.

Russian National Corpus:

The main corpus, which includes texts representing standard Russian, can be subdivided into 3 parts, each of which has its distinguishing features: modern written texts (from the 1950s to the present day), a sub corpus of real-life Russian speech (recordings of oral speech from the same period), and early texts (from the middle of the 18th to the middle of the 20th centuries).

Every text included in the main corpus is subject to meta tagging and morphological tagging. In a small part of the main corpus (currently around 5 million tokens).

http://www.ruscorpora.ru
Automatic Knowledge Control System (AKCS)

Vector Similarity

\[ \text{sim}(A, B) = \cos(\theta) = \frac{A \cdot B}{\|A\|\|B\|} \]

- Calculate the average vector for all words in every sentence/document and use cosine similarity between vectors.
- Calculate weight for foreign words, numbers, and formulas from the text used fingerprint algorithm.
Questions were taken for the central control of knowledge in the discipline "Fundamentals of the development of corporate infocommunication systems". The data set as a whole contains 90 questions and 800 student responses. Responses were evaluated by experts using ratings from 0 to 100 percent.

The results of the designed system are very close to the results of manual verification. In some cases, estimates overlap, while in others, the difference is from 2 to 6%.
Thank you for attention!

Speaker’s contacts:

Naeim Rezaeian
Peoples Friendship University of Russia (RUDN)
Naeim.Rezaeian@hotmail.com