Project Report
On
Automatic Concept Map Generation from Text-based Learning Material

INDIAN INSTITUTE OF TECHNOLOGY, KHARAGPUR

Submitted by:

Laveen Ekka	17CS60R64
Sumit Verma	17CS60R15
Umang Chaturvedi	17CS60R69
Vikash K. Chaurasia	17CS60R51

Under the guidance of:
Prof. Plaban k. .Bhowmick

Contents

Concept Map

A concept map or conceptual diagram is a diagram that depicts suggested relationships between concepts. It is a graphical tool that instructional designers, engineers, technical writers, and others use to organize and structure knowledge.

A concept map typically represents ideas and information as boxes or circles, which it connects with labeled arrows in a downward-branching hierarchical structure. The relationship between concepts can be articulated in linking phrases such as *causes*, *requires*, or *contributes to*

The technique for visualizing these relationships among different concepts is called *concept mapping*. Concept maps have been used to define the ontology of computer systems, for example with the object-role modeling or Unified Modeling Language formalism.

A concept map is a way of representing relationships between ideas, images, or words in the same way that a sentence diagram represents the grammar of a sentence, a road map represents the locations of highways and towns, and a circuit diagram represents the workings of an electrical appliance. In a concept map, each word or phrase connects to another, and links back to the original idea, word, or phrase. Concept maps are a way to develop logical thinking and study skills by revealing connections and helping students see how individual ideas form a larger whole. An example of the use of concept maps is provided in the context of learning about types of fuel.

Concept maps were developed to enhance meaningful learning in the sciences. A well-made concept map grows within a *context frame* defined by an explicit "focus question", while a mind map often has only branches radiating out from a central picture. Some research evidence suggests that the brain stores knowledge as productions (situation-response conditionals) that act on declarative memory content, which is also referred to as chunks or propositions. Because concept maps are constructed to reflect organization of the declarative memory system, they facilitate sense-making and meaningful learning on the part of individuals who make concept maps and those who use them.

Concept maps are widely used in education and business. Uses include:

- Note taking and summarizing gleaning key concepts, their relationships and hierarchy from documents and source materials
- New knowledge creation: e.g., transforming tacit knowledge into an organizational resource, mapping team knowledge
- Institutional knowledge preservation (retention), e.g., eliciting and mapping expert knowledge of employees prior to retirement
- Collaborative knowledge modeling and the transfer of expert knowledge
- Facilitating the creation of shared vision and shared understanding within a team or organization
- Instructional design: concept maps used as Ausubelian "advance organizers" that provide an initial conceptual frame for subsequent information and learning.

- Training: concept maps used as Ausubelian "advanced organizers" to represent the training context and its relationship to their jobs, to the organization's strategic objectives, to training goals.
- Communicating complex ideas and arguments
- Examining the symmetry of complex ideas and arguments and associated terminology
- Detailing the entire structure of an idea, train of thought, or line of argument (with the specific goal of exposing faults, errors, or gaps in one's own reasoning) for the scrutiny of others.
- Enhancing metacognition (learning to learn, and thinking about knowledge)
- Improving language ability
- Assessing learner understanding of learning objectives, concepts, and the relationship among those concepts
- Lexicon development

What is DBpedia ?

DBpedia is a crowd-sourced community effort to extract structured information from Wikipedia and make this information available on the Web. DBpedia allows you to ask sophisticated queries against Wikipedia and link these data to the different datasets on the Web. Furthermore, it might inspire new mechanisms for navigating, linking, and improving the encyclopedia itself.

What is DBpedia Spotlight?

DBpedia Spotlight is a tool for automatically annotating mentions of DBpedia resources in text, providing a solution for linking unstructured information sources to the Linked Open Data cloud through DBpedia. Currently, DBpedia Spotlight contains two approaches: Model and Lucene.
Model was described in the article “Improving Efficiency and Accuracy in Multilingual Entity Extraction” in the Proceedings of the 9th International Conference on Semantic Systems (*I-Semantics 2013*).
Lucene version was described in the article DBpedia Spotlight: Shedding Light on the Web of Documents was published in the Proceedings of the 7th International Conference on Semantic Systems (*I-Semantics 2011*).

Concept Similarity

Calculating the similarity between the concepts is the key in the process of the ontology mapping. First we convert word to vector and then find cosine similarity between them

The cosine of two non-zero vectors can be derived by using the Euclidean dot product formula:

$$\text{similarity} = \cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{||\mathbf{A}|| \, ||\mathbf{B}||} = \frac{\sum\limits_{i=1}^{n} A_i B_i}{\sqrt{\sum\limits_{i=1}^{n} A_i^2} \sqrt{\sum\limits_{i=1}^{n} B_i^2}}$$

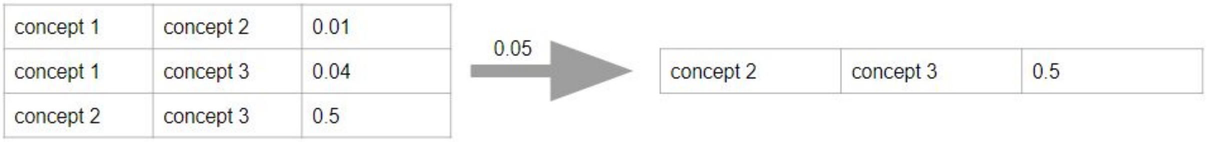
The resulting similarity ranges from −1 meaning exactly opposite, to 1 meaning exactly the same, with 0 indicating orthogonality (decorrelation), and in-between values indicating intermediate similarity or dissimilarity.

For text matching, the attribute vectors *A* and *B* are usually the term frequency vectors of the documents. The cosine similarity can be seen as a method of normalizing document length during comparison.

In the case of information retrieval, the cosine similarity of two documents will range from 0 to 1, since the term frequencies (tf-idf weights) cannot be negative. The angle between two term frequency vectors cannot be greater than 90°.

Filtering Unrelated Pairs

Threshold over the similarity values to keep relations that associates concepts having higher similarity value. You may also use proximity threshold to filter out pairs that are distant apart in document. Constraint such as the concepts have to in the same sentence in order to be related.



concept 1	concept 2	0.01
concept 1	concept 3	0.04
concept 2	concept 3	0.5

concept 2	concept 3	0.5

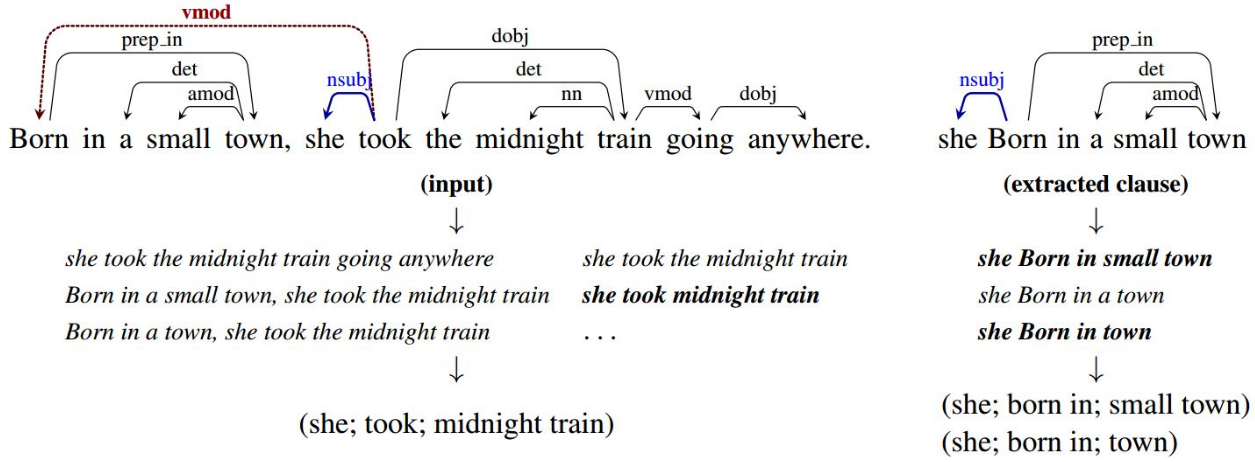
Extracting Relations

Use OpenIE tool to parse the sentences involving a pair and extract <subject, predicate, object> triple. For example. OpenIE output of “Barack Obama was born in Hawaii.” would be (Barack Obama; was born in; Hawaii). Use the predicate name as relation.

OpenIE

Open information extraction (open IE) refers to the extraction of relation tuples, typically binary relations, from plain text. The central difference is that the *schema* for these relations does not need to be specified in advance; typically the relation name is just the text linking two arguments. For example, *Barack Obama was born in Hawaii* would create a triple (Barack Obama; was born in; Hawaii), corresponding to the open domain relation was-born-in(Barack-Obama, Hawaii). This software is a Java implementation of an open IE system as described in the paper:

The system first splits each sentence into a set of entailed clauses. Each clause is then maximally shortened, producing a set of entailed shorter sentence fragments. These fragments are then segmented into OpenIE triples, and output by the system. An illustration of the process is given for an example sentence below:



Born in a small town, she took the midnight train going anywhere.

(input)

↓

she took the midnight train going anywhere

Born in a small town, she took the midnight train

Born in a town, she took the midnight train

she took the midnight train

she took midnight train

. . .

↓

(she; took; midnight train)

she Born in a small town

(extracted clause)

↓

she Born in small town

she Born in a town

she Born in town

↓

(she; born in; small town)
(she; born in; town)

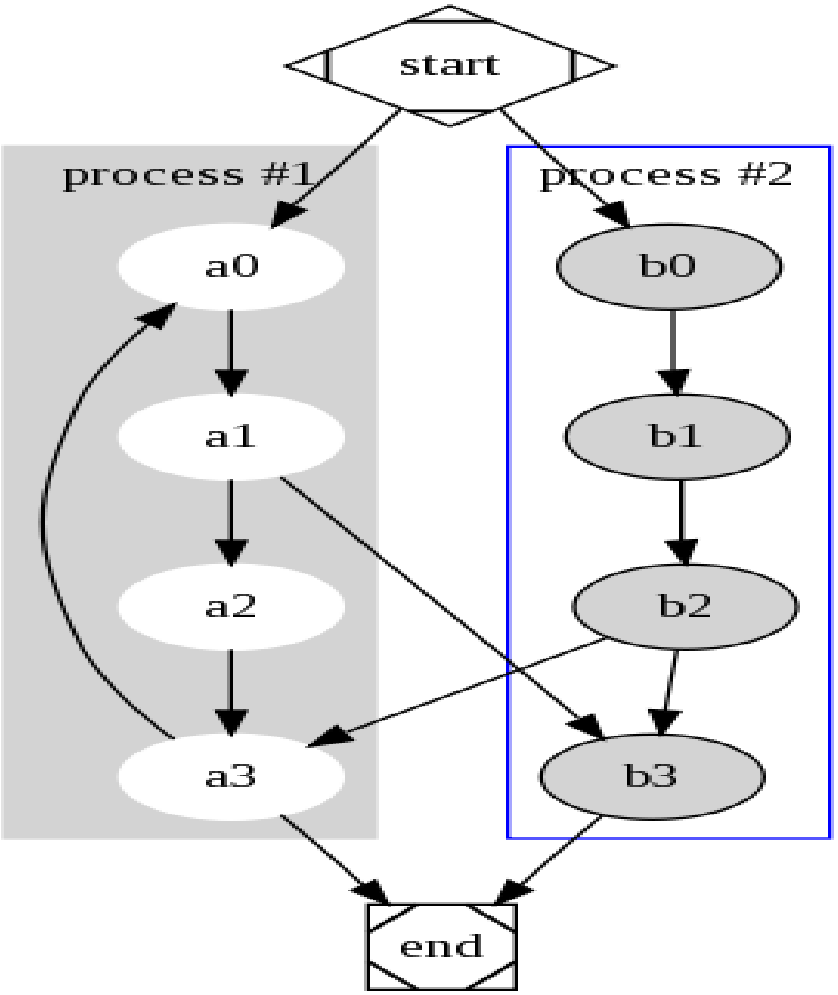
The system was originally written by Gabor Angeli and Melvin Johnson Premkumar. It requires Java 1.8+ to be installed, and generally requires around 50MB of memory in addition to the memory used by the part of speech tagger and dependency parser (and optional named entity recognizer). We recommend running java with around 1gb of memory (2gb if using NER) to be safe (i.e., java -mx1g).
The system is licensed under the GNU General Public License (v2 or later). Source is included. The package includes components for command-line invocation, and a Java API. The code is dual licensed (in a similar manner to MySQL, etc.). Open source licensing is under the *full* GPL, which allows many free uses. For distributors of proprietary software, commercial licensing is available. If you don't need a commercial license, but would like to support maintenance of these tools, we welcome gift funding.

Graph generation

Finally, the generated concept graph has to be stored against the input document and can be visualized. There are several algorithms for graph/network visualization. We are using graphviz.

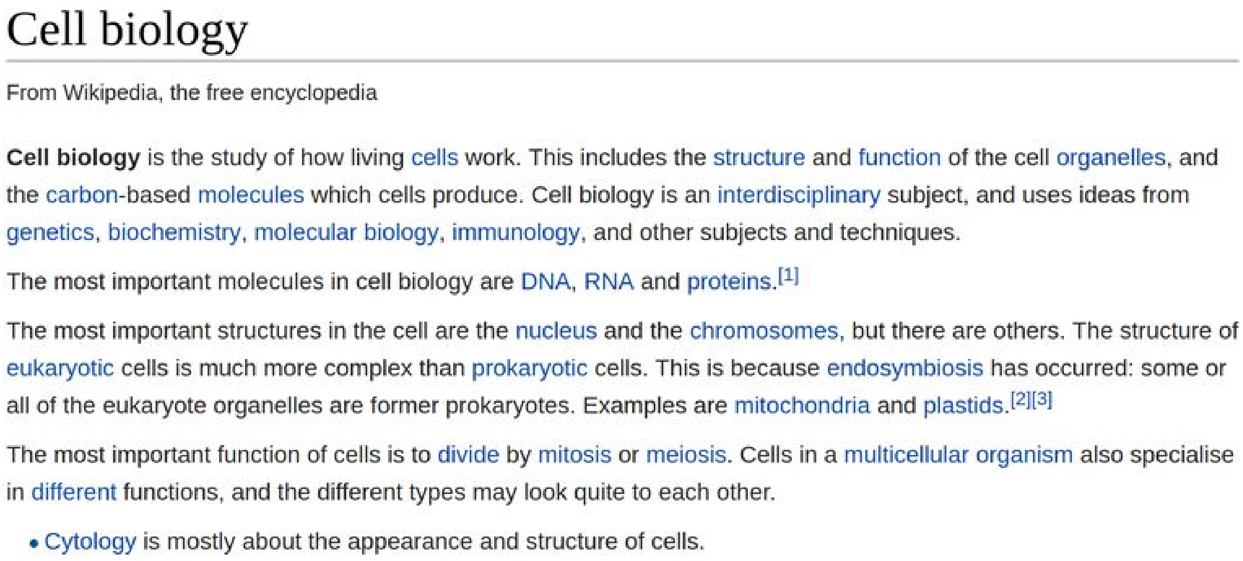
What is Graphviz?

Graphviz is open source graph visualization software. Graph visualization is a way of representing structural information as diagrams of abstract graphs and networks. It has important applications in networking, bioinformatics, software engineering, database and web design, machine learning, and in visual interfaces for other technical domains.



Example

For the below input text to our system



Cell biology

From Wikipedia, the free encyclopedia

Cell biology is the study of how living cells work. This includes the structure and function of the cell organelles, and the carbon-based molecules which cells produce. Cell biology is an interdisciplinary subject, and uses ideas from genetics, biochemistry, molecular biology, immunology, and other subjects and techniques.

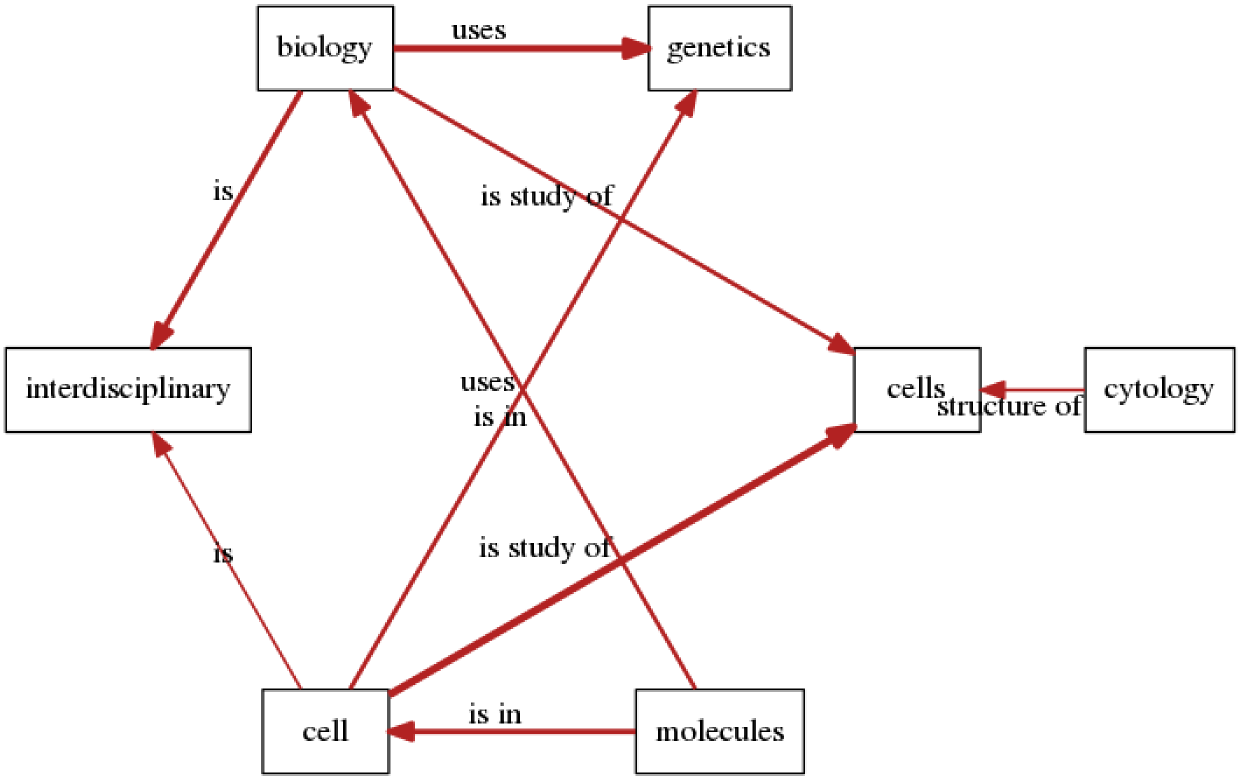
The most important molecules in cell biology are DNA, RNA and proteins.[1]

The most important structures in the cell are the nucleus and the chromosomes, but there are others. The structure of eukaryotic cells is much more complex than prokaryotic cells. This is because endosymbiosis has occurred: some or all of the eukaryote organelles are former prokaryotes. Examples are mitochondria and plastids.[2][3]

The most important function of cells is to divide by mitosis or meiosis. Cells in a multicellular organism also specialise in different functions, and the different types may look quite to each other.

- Cytology is mostly about the appearance and structure of cells.

The concept map generated is as following:



Evaluation Methods

Evaluation can be performed by both the human based method and the system based methods. Concept maps are not unique so we need some measure to evaluate the system generated concept map. We can use the majority voting method by experts. For a given concept map, we can ask experts to categorize that into “good”, “average”, “poor” categories. Based on the category which has got the maximum votes for a particular text, we can evaluate the concept map.

The system based evaluation approach can be use of summary evaluation techniques. We can generate summary from the concept map and we can evaluate the summary by rouge scores. We can report rouge score for a generated concept map. Higher the rouge score indicates better the concept map. This is a kind of reverse engineering method to get the performance metrics of the developed system.