Avdonina Marina, Rudneva Maria, Valeeva Nailia, Zhabo Natallia. Training of the translation of new environmental terms in Russian, English and French (study-synthesis) in the professional vocabulary. In: EDULEARN 16. 8th International conference on education and new learning technologies. 2016. PP. 8421–8427.

Chernysheva Irina, Zhabo Nathalia, Avdonina Marina. 2019. A prior discussion in the native language of the situation and themes of the text as an essential component of translational exercises system. Foreign Languages for Special Purposes. Yerevan: Yerevan State University Press. № 7 (16). 69–82.

Gouadec, Daniel. 1989. Comprendre, évaluer, prévenir. TTR (2:2) L'erreur en traduction, 35-54.

Kleiber, Georges. 2008. Petit essai pour montrer que la polysémie n'est pas un sens interdit. In Durand J. Habert B., Laks B. (éds), Actes du CMLF, 87-100.

Ngoc Quan Tran. 2017. Étude des titres de presse : classement syntaxique, valeurs sémantiques et pragmatiques. Linguistique. https://dumas.ccsd.cnrs.fr/dumas-01558210/document

Teaching Computers to Read, Understand, and Write Human Languages

Ben Choi[1], Andrey Timofeyev[2], Anna Bobunova[3]

[1]Associate Professor,
Louisiana Tech University, USA

[2]Lecturer,
Louisiana Tech University, USA

3Senior Lecturer
RUDN University, Russia

Abstract

This paper presents a contemporary linguistic challenge: how to teach computers to read, understand, and write human languages. Researchers in linguistics and artificial intelligence need to work together to address this challenge. This paper outlines an attempt to address this challenge by creating an artificial intelligent system that can compose new English sentences to summarize documents. To process the documents conceptually to create abstractive summaries, the system makes use of one of the world's largest knowledgebase and one of the most powerful inference engines. The resultant AI system first uses natural language processing techniques to extracts syntactic structure of the documents and then maps the words of the sentences and their parts of speech into related concepts in the knowledgebase. It then uses the inference engine to generalize and fuse concepts to form more abstract concepts. The system then composes new sentences based on the key concepts by linking subject concepts with their related predicate concepts. The system has been implemented and tested. The test results showed that the system can create new sentences that include abstracted concepts not explicitly mentioned in the original documents and that contain information synthesized from different parts of the documents to compose a summary.

Key words: *Computational Linguistics, Natural Language Processing, Artificial Intelligence, Text Summarization, Knowledge Engineering.*

INTRODUCTION

This paper examines a contemporary linguistic challenge: how to teach computers to read, understand, and write human languages. Researchers in Computational Linguistics (including researchers in linguistics and artificial intelligence) have been working together to address this challenge that is also referred to as Natural Language Processing. However, this challenge involves not only Natural Language Processing of linguistic data, but also requires Knowledge Engineering (Choi 2021), which makes use of knowledge bases and inference engines to process knowledge. The ability to read, understand, and write human languages requires not only processing the given text data, but also requires commonsense knowledge. The additional knowledge required are encoded into computer programs and databases. Programs for parsing, encoded with the knowledge of grammar, read and analyze the given text data syntactically to produce parts of speech. Knowledgebases, encoded with the ontology of human knowledge, provide concepts that relate words semantically to their meanings. Inference engines, encoded with the knowledge of reasoning, deduce and generate new concepts from the contents of the given text documents and thus give rise to “understand”. Programs for writing, encoded with the knowledge of grammar, utilize the resultant new concepts and the relationships between concepts to compose new sentences.

This paper outlines an attempt to address the challenge by creating an artificial intelligent system that can compose new English sentences to summarize documents (Choi & Huang 2010). We describe how to build knowledge based automatic summarization system that can create abstractive summaries of the given documents by generalizing new concepts, deriving main topics, and composing new sentences. The system processes text data on documents and webpages and utilizes knowledgebase and inference engine to produce an abstractive summary. It generates summaries by composing new sentences based on the semantics derived from the text.

To process the documents conceptually to create abstractive summaries, the system makes use of one of the world’s largest knowledgebase and one of the most powerful inference engines. The system uses both the syntactic structure provided by the given documents and the commonsense knowledge provided by the knowledgebase. It performs deep syntactic analysis by using capabilities of advanced natural language processing techniques. It uses Cyc development platform as a source of background knowledge (cyc.com). The Cyc development platform consists of the world’s largest ontology of commonsense knowledge and a reasoning engine that allows information comprehension and abstraction. In addition, Cyc ontology serves as a backbone for semantic analysis, knowledge generalization, and natural language generation.

We attempt to develop a system that can compose new sentences for the summary. Our system generalizes new abstract concepts based on the knowledge derived from the text. It automatically detects main topics described in the text. Moreover, it composes new English sentences for some of the most significant concepts. The created sentences form an abstractive summary, combining concepts from different parts of the input text.

The system conducts summarization process in three principal stages: knowledge acquisition, knowledge discovery, and knowledge representation. The knowledge acquisition stage derives syntactic structure of each sentence of the input document and maps words and their relations into Cyc knowledgebase. Next, the knowledge discovery stage generalizes concepts upward in the Cyc ontology and detects main topics covered in the text. Finally, the knowledge representation stage composes new sentences for some of the most significant concepts defined in main topics. The syntactic structure of the newly created sentences follows an enhanced subject-predicate-object model, where adjective and adverb modifiers are used to produce more complex and informative sentences.

We have implemented our proposed system that was tested on various documents and webpages. The test results show that our system is capable of identifying key concepts and discovering main topics comprised in the original text, generalizing new concept not explicitly mentioned in the text, and creating new sentences that contain information synthesized from various parts of the text. The newly created sentences have complex syntactic structures that enhance subject-predicate-object triplets with adjective and adverb modifiers. For example, the sentence “Colored grapefruit being sweet edible fruit” was automatically generated by the system analyzing encyclopedia articles describing grapefruits. Here, the subject concept “grapefruit” is modified by the adjective concept “colored” that was not explicitly mentioned in the text and the object concept “edible fruit” is modified by the adjective concept “sweet”.

The sentence was created as the result of linked key concepts. The linked concepts are then mapped back to words to form the sentence. As we can see from the above example, the created sentence sound like made by a machine. Human writers might use the word “is” instead of “being” in the example sentence. In short, although our system can generate new abstractive sentences, there is much more research potential to further develop such a knowledge-based system to compose new sentences as summary.

KNOWLEDGE-BASED ABSTRACTIVE SUMMARIZATION

Our abstractive knowledge-based summarization system attempts to bring the machines one-step closer to the comprehension of the knowledge comprised in the text. The system performs text summarization in three principal steps (Figure 1): the knowledge acquisition, the knowledge discovery, and the knowledge representation (Timofeyev & Choi 2019).

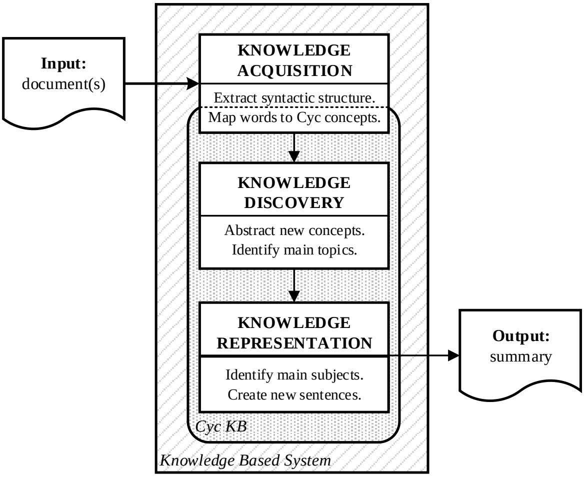

Figure 1. System workflow

During the knowledge acquisition step, the algorithm receives text documents as an input, performs syntactic analysis, and maps the words with their syntactic relationships into the Cyc knowledgebase. During the knowledge discovery step, the system performs a generalization of new concepts by propagating the concepts that were mapped into Cyc knowledgebase by the knowledge acquisition step. It also performs the task of the identification of the main topics of the text based on the mapped and generalized concepts. Finally, during the knowledge representation step, the system generates new sentences using knowledge derived from the input text documents and the capabilities of the Cyc inference engine.

Knowledge Acquisition

The knowledge acquisition step consists of two subprocesses. The first subprocess extracts the syntactic structures from the given documents (Figure 2). This subprocess serves as a data preprocessing and transformation step. It normalizes raw text data and transforms it into syntactic representation.

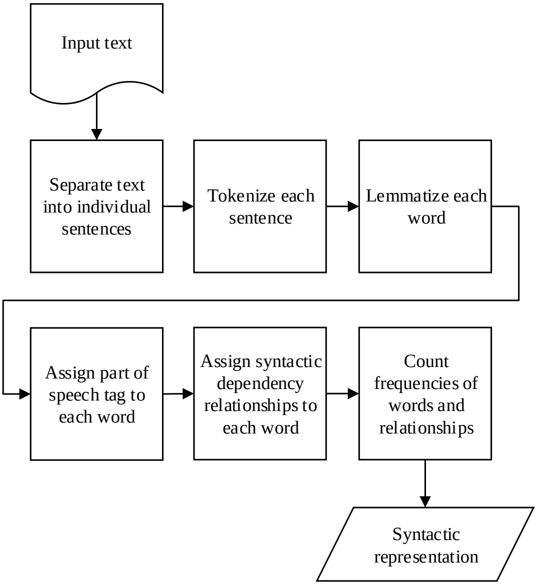
Input text
Separate text into individual sentences
Tokenize each sentence
Lemmatize each word
Assign part of speech tag to each word
Assign syntactic dependency relationships to each word
Count frequencies of words and relationships
Syntactic representation

Figure 2. Syntactic structure extraction

The second subprocess maps words from syntactic representation of the text to Cyc concepts (Fig. 3). Mapped Cyc concepts are utilized for reasoning during subsequent steps of the algorithm.

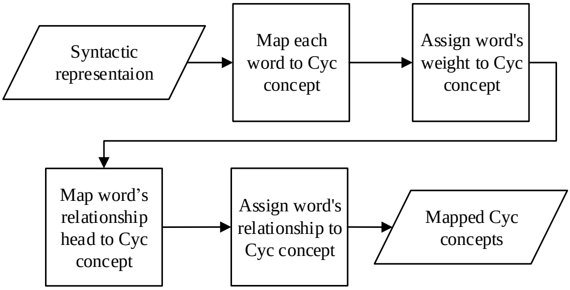
Syntactic representaion
Map each word to Cyc concept
Assign word's weight to Cyc concept
Map word’s relationship head to Cyc concept
Assign word's relationship to Cyc concept
Mapped Cyc concepts

Fig. 3. Mapping words to Cyc concepts

Knowledge Discovery

The knowledge discovery step performs two subprocesses: it abstracts new concepts and identifies main topics described in the input text. New concepts abstraction subprocess (Fig. 4) generalizes the information derived from the text. It finds the ancestors of mapped Cyc concepts and assigns the descendants’ propagated weight and syntactic dependency relationships to the ancestors (Choi & Huang 2009). It is an important part of the abstractive summarization process as it allows deriving concepts that are not explicitly mentioned in the input text. For example, concepts like “cat,” “tiger,” “jaguar,” and “lion” are generalized into more abstract “feline” concept.

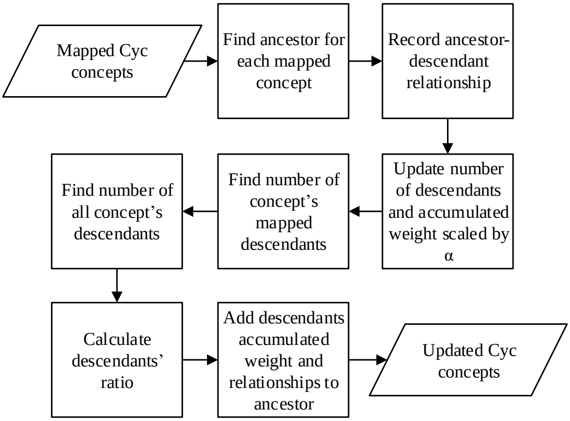
Mapped Cyc concepts
Find ancestor for each mapped concept
Record ancestor-descendant relationship
Update number of descendants and accumulated weight scaled by α
Find number of concept’s mapped descendants
Find number of all concept’s descendants
Calculate descendants’ ratio
Add descendants accumulated weight and relationships to ancestor
Updated Cyc concepts

Fig. 4. New concepts abstraction

The main topics identification subprocess (Fig. 5) detects topics described in the text with an assumption that they are represented by the most frequently used micro theories. Micro theories form the basis of the knowledge organization in Cyc ontology being the clusters of Cyc concepts and facts, typically representing one specific domain of knowledge. For example, #$BiologyMt is a micro theory containing biological knowledge, and #$MathMt is a micro theory containing concepts and facts describing the field of mathematics. Each Cyc concept is defined within a micro theory.

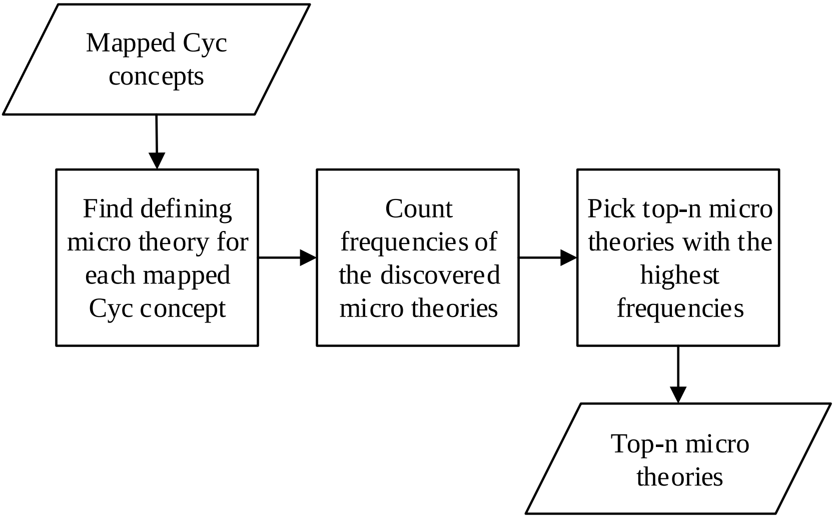
Mapped Cyc concepts
Find defining micro theory for each mapped Cyc concept
Count frequencies of the discovered micro theories
Pick top-n micro theories with the highest frequencies
Top-n micro theories

Fig. 5. Main topics identification

Knowledge Representation

The knowledge representation utilizes powerful capabilities of the Cyc inference engine to generate new sentences based on the information discovered during knowledge acquisition and knowledge discovery steps. This step uses mapped and generalized Cyc concepts, their syntactic dependency relationships, and the most frequent micro theories as inputs. The knowledge representation step consists of two subprocesses: candidate subject discovery and new sentence generation. The candidate subject discovery subprocess (Fig. 6) identifies significant subject concepts out of all the mapped and generalized Cyc concepts.

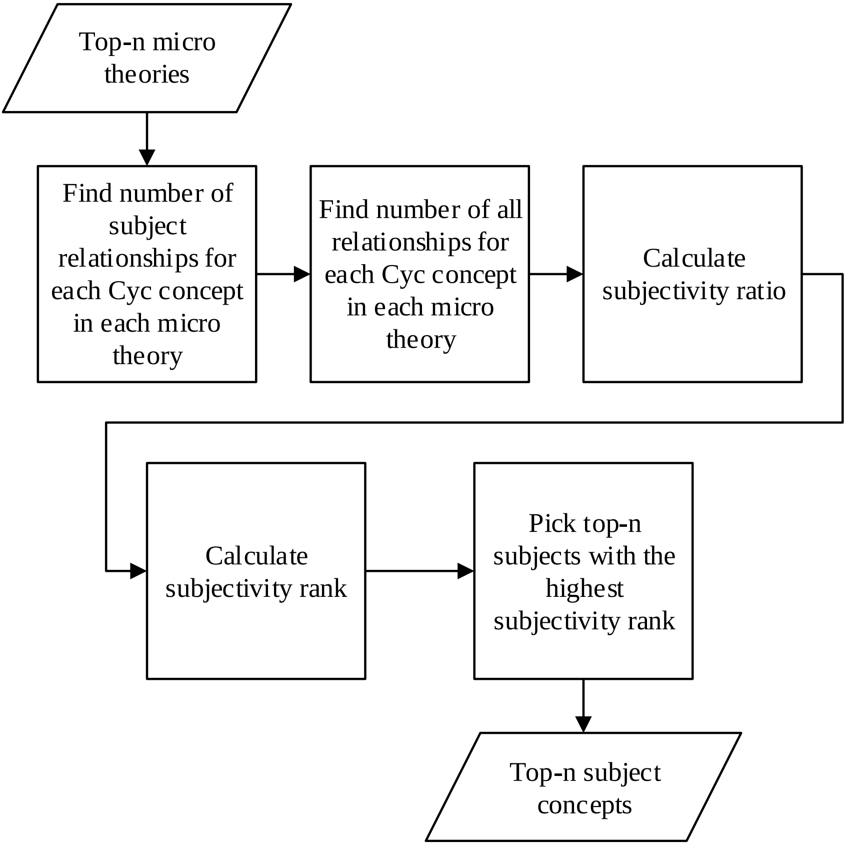
Top-n micro theories
Find number of subject relationships for each Cyc concept in each micro theory
Find number of all relationships for each Cyc concept in each micro theory
Calculate subjectivity ratio
Calculate subjectivity rank
Pick top-n subjects with the highest subjectivity rank
Top-n subject concepts

Fig. 6. Candidate subject discovery

The new sentences generation subprocess (Fig. 7) composes new sentences for each of the identified candidate subject concepts (Choi & Huang 2010). The generated sentences serve as a final summary of the input text.

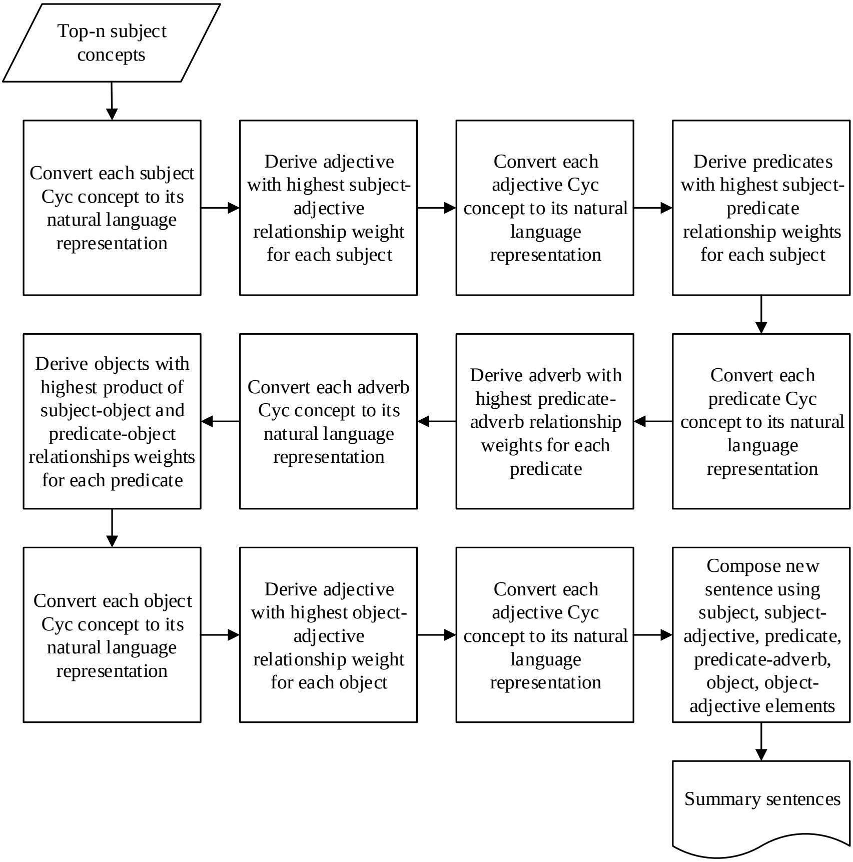
Top-n subject concepts
Convert each subject Cyc concept to its natural language representation
Derive adjective with highest subject-adjective relationship weight for each subject
Convert each adjective Cyc concept to its natural language representation
Derive predicates with highest subject-predicate relationship weights for each subject
Convert each predicate Cyc concept to its natural language representation
Derive adverb with highest predicate-adverb relationship weights for each predicate
Convert each adverb Cyc concept to its natural language representation
Derive objects with highest product of subject-object and predicate-object relationships weights for each predicate
Convert each object Cyc concept to its natural language representation
Derive adjective with highest object-adjective relationship weight for each object
Convert each adjective Cyc concept to its natural language representation
Compose new sentence using subject, subject-adjective, predicate, predicate-adverb, object, object-adjective elements
Summary sentences

Fig. 7. New sentence generation

CONCLUSIONS AND FUTURE WORK

The system has been implemented and tested (Timofeyev & Choi 2018). The test results showed that the system can create new sentences that include abstracted concepts not explicitly mentioned in the original documents and that contain information synthesized from different parts of the documents to compose a summary. The task of producing an abstractive summary of a given text is considered challenging for humans and even more so for machines. Employing the semantic features and the syntactic structure of the text together with the world’s largest knowledge base shows great potential in creating abstractive summaries. Although our system can generate new abstractive sentences, there is much more research potential to further develop such a knowledge-based system to compose new sentences as summary.

Future research potential includes enhancing the domain knowledge since the semantic knowledge and reasoning are limited to the functionality and performance of the underlining commonsense knowledgebase. Our system is currently as knowledgeable as the capabilities of the Cyc knowledgebase that is currently the largest ontology of commonsense knowledge. For future improvement, a system could use the information derived from the whole World Wide Web as domain knowledge. This would possess challenging research questions such as information inconsistency and sense disambiguation. In addition, a robust inference engine would be required to process the information correctly and in a timely fashion.

Another potential research is to improve the process and the structure of composing new sentences. Our system currently uses subject-predicate-object triplets enhanced by adjective and adverb modifiers. It does not yet resemble the structure of the sentences created by human. The structure of newly created sentences can be improved by using a more sophisticated representation of the syntactic structure of the sentence, such as graph representation. Moreover, another future research direction is to compose several connected sentences to form a coherent abstract. Currently, the sentences created by our system are not directly connected to each other. One possible enhancement is by representing the whole document as a graph of connected concepts with various relationships among them and then creating new sentences based on these relationships. Much more research is needed for a machine to create a coherent abstract to summarize documents.

References

Choi, Ben. 2021. “Knowledge Engineering the Web,” *International Journal of Machine Learning and Computing* vol. 11, no. 1, pp. 68-76.

Choi, Ben & Xiaomei Huang. 2009. “Web Page Summarization by Using Concept Hierarchies,” *International Conference on Agents and Artificial Intelligence (ICAART 2009)*, pp.281-286.

Choi, Ben & Xiaomei Huang. 2010. “Creating New Sentences to Summarize Documents,” *The 10th IASTED International Conference on Artificial Intelligence and Application (AIA 2010)*, pp. 458-463.

Timofeyev, Andrey & Ben Choi. 2018. “Building a Knowledge Based Summarization System for Text Data Mining,” *Machine Learning and Knowledge Extraction*, pp. 118-133.

Timofeyev, Andrey & Ben Choi. 2019. “Knowledge Based System for Composing Sentences to Summarize Documents,” *Knowledge Discovery, Knowledge Engineering and Knowledge Management*, pp. 164-183.

The Influence of the Persian Civilization on the Vocabulary of the Chinese Language during the Great Silk Road

Kamilla A. Dana

PhD lecturer of the Department of Chinese Philology
Tashkent State University of Oriental Studies, Uzbekistan

The main way to enrich the vocabulary of the Chinese language throughout its development was the formation of new words based on the existing building material. However, there is another source of replenishment of the vocabulary of the language. These are foreign borrowings, 外来词, literally "words that came from outside" (Semenas, 2007). The influence of one language on another is most clearly expressed not in phonetics or grammar, but in vocabulary, in borrowing one language from another. New realities are usually borrowed along with their designations.