NATURAL LANGUAGE PROCESSING FOR THE CASE OF THE LEXICAL TEXT ADAPTATION AND GENERATING GRAMMAR EXERCISES

Baglaeva E.A.
Scientific supervisor: Tsapko S.G., Candidate of Engineering Sciences, Associate Professor
Language advisor: Shepetovsky D.V., senior teacher
Tomsk Polytechnic University, 30, Lenin Avenue, Tomsk, Russia, 634050
E-mail: eab14@tpu.ru

Natural language processing (NLP) is a field of computer science, artificial intelligence, and linguistics concerned with the interactions between computers and human (natural) languages.

As such, NLP is related to the area of human–computer interaction. Many challenges in NLP involve natural language understanding, that is, enabling computers to derive meaning from human or natural language input, and others involve natural language generation. [1]

Lexical adaptation of news articles and automated generation of grammar exercises, based on these and other texts, will give us possibility to create an unlimited number of new individual tasks and tests for second language learners.

While bringing forth this problem, we construct a model of the solution.

During the course of construction of the model we divide knowledge about language into

1) General (basic rules of grammar) ;
2) Individual (some exceptions to the rule of lexemes).

Basic grammar rules allow us to formulate a parsing algorithm, applicable to a particular grammar (for example, English).

According to the Chomsky hierarchy of formal grammars [2], English language has a lot in common with context-free grammars, which simplifies its handling by a particular linear algorithm. However, the question whether the English is context-free grammar is still not resolved.

While constructing the model a question arises whether the interpretation obtained by the system conveys the whole morphological, syntactic and semantic meaning, and whether such interpretation is necessary.

In most cases such detailed interpretation of words is not necessary.

Interpretation is necessary in specific cases:

1) to distinguish the word A from word B;
2) to implement implicative and synonymous substitution;
3) to generate abstracts;
4) to search for knowledge in a semantic network;
5) to provide question-answer communication with the system.

Requirements for word interpretation:

1) Determine the lexical meaning through simple values;
2) Lexical meanings B and C should be sufficient to determine the word A;
3) Uniform interpretation of A and B, despite the difference in the external forms of A and B. [3]

In the interpretation, first, typical situation with a particular word is analyzed, and then a deeper analysis starts with the words associated with the particular word lexically, semantically or morphologically.

For the case of the text adaptation in question, implementation and use of a slightly modified dictionary of synonyms and synonymous expressions is sufficient for interpretation words.

Construction of the model requires classification of the knowledge on the basis of other criteria:

1) Active (the system uses in its work);
2) Passive (information stored in the system but not used by it, including the self-referential information that the system gives to its user).

We define meta-knowledge as information and method of its organization.

Static meta-knowledge is used by the system, but never modified.

Dynamic meta-knowledge may be changed while the system is functioning.

A linguistic processor converts data from one logical level to another.

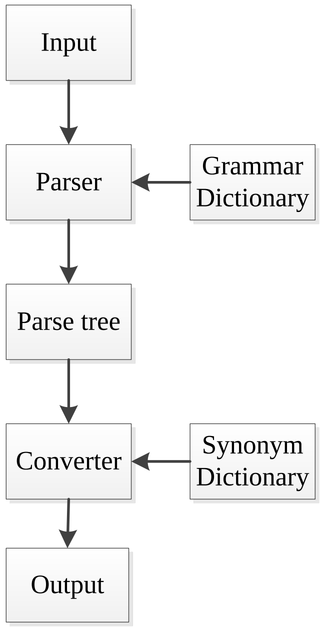

Fig. 1. Structure of the linguistic processor

When constructing linguistic processor it is necessary to determine:

1) Formal languages employed to represent linguistic levels;
2) The formal concept of sentence structure for these levels;
3) The rules for cross-converting arrays of structures;
4) Morphological, collocational and semantic dictionaries.

The linguistic processor can solve the problem of automatic database replenishment directly from the texts.

Principles of the linguistic processor construction:

1) Linguistic knowledge is independent from the language;
2) Grammar dictionaries are independent of algorithms;
3) Linguistic knowledge is independent from a subject area;
4) Linguistic knowledge is independent from the nature of the problem being solved. [4]

Parsing or **syntactic analysis** is the process of analyzing a string of symbols, either in natural language or in computer languages, according to the rules of a formal grammar.

A **parser** is a software component that takes input data (frequently text) and builds a data structure – often some kind of parse tree, abstract syntax tree or other hierarchical structure – giving a structural representation of the input, checking for correct syntax in the process. [5]

There are two parsing strategies:

1) To identify the appropriate case frame and activate its case-markers and filter-patterns to deal with the rest of the input utterance (case-oriented strategy);

2) To recognize individual constituent case filter and markers, including the verb, noun phrases in the role of case filters, and prepositions in the role of case markers (linear pattern-matching strategy). [6]

Translation from one language to another requires the use of transfer rules, which are easy to establish and quite satisfactory if a literal translation between two languages (usually closely-related ones) is feasible. An adaptation of the English text to a given learner’s level is similar to such task, technically it is translation between different forms of English.

Summing up these ideas is an **algorithm of text adaptation** given below.

Input: Article in English.

Output: Article adapted at a certain learner’s skill level in English.

Step 1. Read the sentence, split into words and parse.

Step 2. Using the parse tree, look up for meaningful parts in the dictionary.

Step 3. Replace all the words and expression above the given skill level with synonyms words or expressions belonging to the required skill level.

Step 4. Repeat from **Step 1**, until the end of the text is reached, then stop.

Basic data for generating grammar exercises should include the following components:

1) A lexicon that describes the vocabulary to be used in the generated texts;

2) A discursive component which indicates what information should appear in the text and its order of appearance;

3) A syntax which allows texts to be created from the lexicon and the discursive component. [7]

Algorithm for generating grammar tasks is almost the same as text adaptation.

Input: Sentence in English.

Output: Several variants of the original sentence.

Step 1. Read the sentence, build a parse tree.

Step 2. Find the semantic part, which exists in the lexicon.

Step 3. Pick up several similar, but incorrect expressions from a specific lexicon of grammatically incorrect units.

Step 4. Make sentences with the chosen words.

Step 5. Make incorrect sentences by swapping some parts of the parse tree (for example, subject and predicate in the interrogative sentence).

References

1. Natural language processing. // Wikipedia [Online service] URL: http://en.wikipedia.org/wiki/Natural_language_processing (Retrieved 09.02.14)
2. Chomsky hierarchy. // Wikipedia [Online service] URL: http://en.wikipedia.org/wiki/Chomsky_hierarchy (Retrieved 09.02.14)
3. Popov E.V. Communication with the computer in natural language. – M.: Chief Editorial Board of physical and mathematical literature, 1982. – 360 p.
4. Linguistic processor for complex information systems / Y.D. Apresian, I.M. Boguslavskiy, L.L. Iomdin, etc. – M.: Science, 1992. - 256 p.
5. Parsing // Wikipedia [Online service] URL: http://en.wikipedia.org/wiki/Parsing (Retrieved 09.02.14)
6. Bolc L. Natural language parsing systems. – Springer-Verlag, Berlin, 1987. – 367 p.
7. Patten, Terry. Systemic text generation as problem solving – Cambridge University Press, 1988. – 214 p