Features of software development for natural language processing

Tukeev U. , Zhumanov Zh. , Rakhimova D.

1. Approaches to and methods for solving of natural language processing problems

Natural Language processing (NLP) is a field of computer science concerned with the interactions between computers and human (natural) languages. NLP also has many common areas with linguistics.

The following is a incomplete list of tasks in NLP:

- Machine translation: automatically translate text from one human language to another.
- Part-of-speech tagging: determine the part of speech for each word in a sentence.
- Question answering: determine an answer to a question ask in natural language.
- Relationship extraction: identify the relationships among named entities in a text.
- Topic segmentation and recognition: separate a text into segments each of which is devoted to a topic, and identify what topic a segment is devoted to.
- Word sense disambiguation: determine a meaning of word with multiple meanings which suits given context.

Some of these tasks have a very close relation to semantics (meaning).

Machine translation methods

Machine translation (MT) is a subfield of computational linguistics, which studies the use of computer software to translate text or speech from one natural language into another. At a basic level, the MT performs a simple replacement of words from one natural language into words of another. Use of more complex methods makes it possible to attempt a more complicated translation, allowing better handling of differences in linguistic typology, phrase recognition and translation of idioms.

The translation process can be described as:

- Decoding of the source text's meaning.
- Recoding of that meaning in the target language.

Behind this seemingly simple procedure lies a complex cognitive activity. To decode the meaning of the original text as a whole, the translator must interpret and analyze all the features of the text - a process that requires a deep knowledge of grammar, semantics, syntax, idioms, etc. the source language and the culture of its speakers. Translator needed such as in-depth knowledge in the target language for the conversion of meaning.

Therein lies the difficulty of machine translation: how to program a computer so he could "understand" the text as it makes people and "create" a new text in the target language, which would be written like a man.

There are several approaches to this problem.

Rule-based MT

The rule-based machine translation consist of several approaches: transfer-based machine translation, interlingual machine translation and dictionary-based machine translation.

Interlingual MT

Interlingual machine translation is one of the classic approaches to machine translation. In this approach, the source text is transformed into an interlingua, an abstract language-independent representation. The target language is then generated from the interlingua.

Advantages of interlingual approach are:

- it requires fewer components in order to relate each source language to each target

language;
- it takes fewer components to add a new language;
- it supports paraphrases of the input in the original language;
- it allows both the analyzers and generators to be written by monolingual system developers;
- it handles languages that are very different from each other.

Disadvantage of interlingual approach is that the definition of an interlingua is difficult and maybe even impossible for a wider domain.

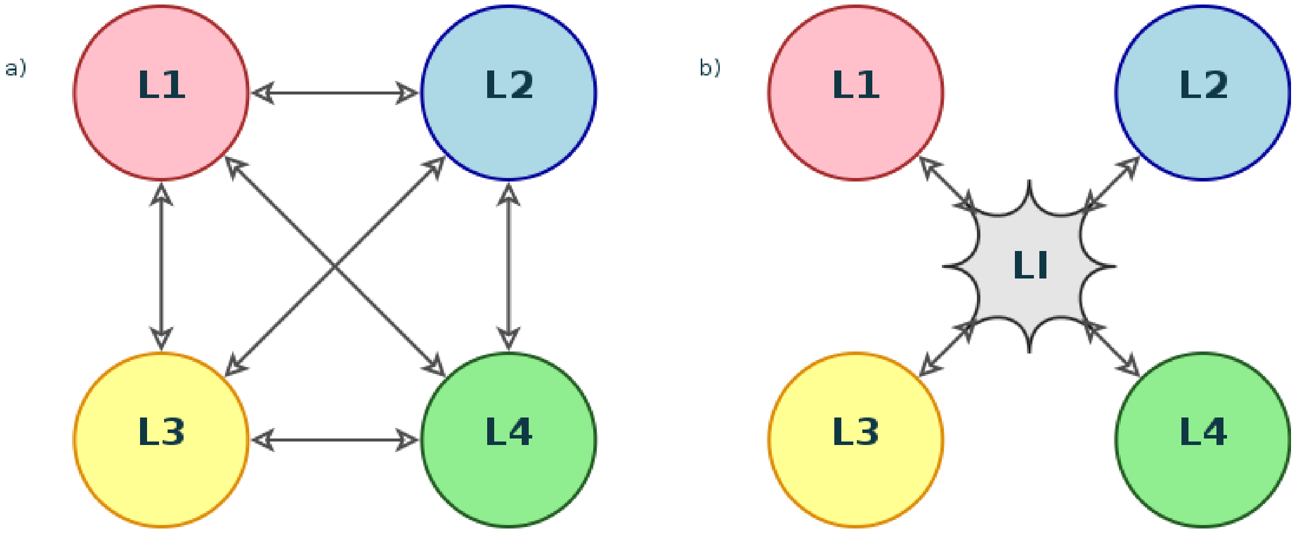

Figure 1.1 a) Translation graph for direct transfer-based machine translation (4 languages, 12 modules required); b) Translation graph for using a bridge language (4 languages, 8 translation modules required).

Sometimes two interlinguas can be used in translation. It is possible that one of the two covers more of the source language's characteristics, and the other covers more of the target language's characteristics. The translation process is shown in the next picture.

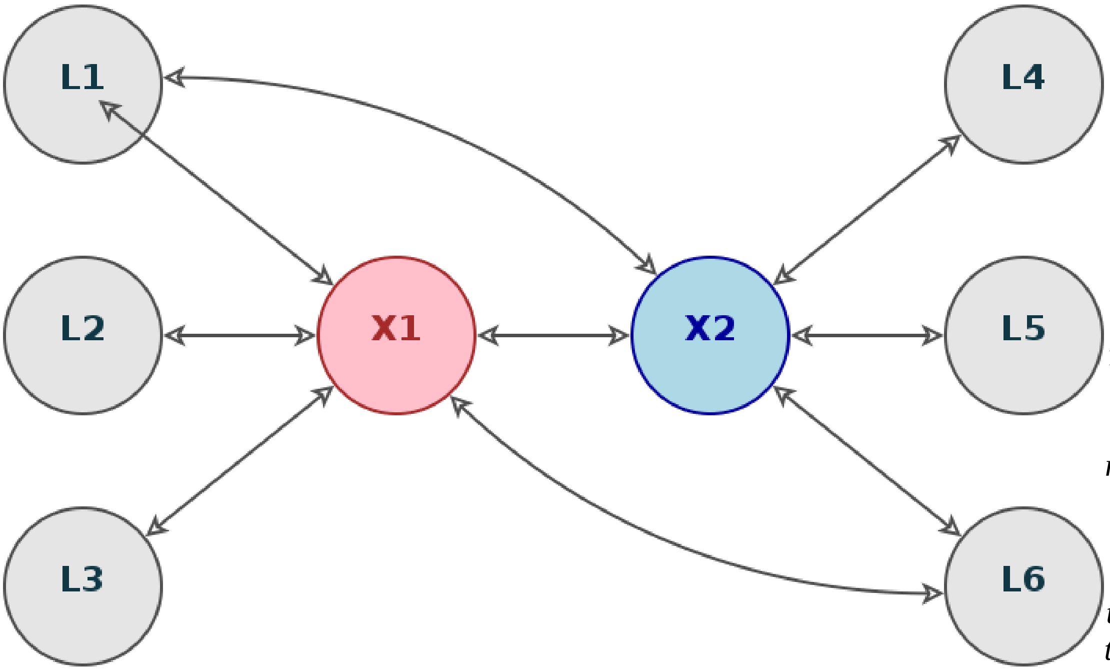

Figure 1.2: Translation using two interlinguas.

Transfer-based machine translation

Transfer-based machine translation is based on the idea of interlingua and is currently one of the most widely used methods of machine translation.

Main idea of transfer-based machine translation is next: it is necessary to have an intermediate representation of the original sentence in order to generate the correct translation. In transfer-based MT intermediate representation has some dependences on the language pair involved. The way in which transfer-based machine translation systems work is simple: it applies sets of linguistic rules which are defined as correspondences between the structure of the source language and that of the target language. The first step includes analyzing the input text for morphology and syntax to create an internal representation. The translation is generated from this representation using bilingual dictionaries and grammatical rules.

Quality of translation using this approach depends on the language pair it is applied to.

Statistical approach

Statistical machine translation (SMT) is an approach to machine translation where translations are generated on the basis of statistical models whose parameters are derived from the analysis of bilingual text corpora.

The idea behind statistical machine translation comes from information theory. A document is translated according to the probability distribution p(e | f) that a string e in the target language is the correct translation of a string f in the source language. The problem of modeling the probability distribution $p(e \mid f)$ has been approached in several ways. One approach is to apply Bayes Theorem, $p(e|f) \propto p(f|e)p(e)$, where the translation model $p(f \mid e)$ is the probability that the source string is the translation of the target string, and the language model $p(e)$ is the probability of seeing that target language string. This decomposition splits the problem into two subproblems. Finding the best translation $\tilde{e}$ is done by picking up the one that gives the highest probability:

$$\tilde{e} = arg \max_{e \in e^*} p(e|f) = arg \max_{e \in e^*} p(f|e)p(e).$$

Texts are typically translated sentence by sentence. Language models are usually approximated by smoothed n-gram models, and similar approaches have been applied to translation models, but there is some complexity due to different sentence lengths and word orders in the languages. The statistical translation models were initially word based, but significant advances were made with the introduction of phrase based models.

Word-based translation

In word-based translation, the fundamental unit of translation is a word in some natural language. Typically, the number of words in translated sentences are different, because of compound words, morphology and idioms. The ratio of the lengths of sequences of translated words is called fertility, which tells how many foreign words each native word produces. Necessarily it is assumed by information theory that each covers the same concept.

Simple word-based translation can't translate between languages with different fertility. Word-based translation systems can relatively simply be made to cope with high fertility, but they could map a single word to multiple words, but not the other way about.

The word-based translation is not widely used today, phrase-based systems are more common. The alignments are used to extract phrases or deduce syntax rules. Matching words in bi-text is still a problem actively discussed in the community.

Phrase-based translation

The aim of phrase-based translation is to reduce the restrictions of word-based translation by translating whole sequences of words, where the lengths may differ. The sequences of words are called blocks or phrases. Usually they are not linguistic phrases but phrases found using statistical methods from corpora. Restricting the phrases to linguistic phrases decreases the quality of translation.

Example-based MT

Example-based machine translation (EBMT) approach to machine translation uses of a bilingual corpus with parallel texts as its main knowledge base at run-time.

At the foundation of example-based machine translation is the idea of translation by analogy. Example-based machine translation systems are trained from bilingual parallel corpora, which contain sentence pairs like the example shown in the table. Sentence pairs contain sentences in one language with their translations into another. So called “minimal pair” consists of the sentences that vary by just one element. These sentences make it simple to learn translations of subsentential units. Composing these units can be used to produce translations in the future.

EBMT is best suited for sub-language phenomena like phrasal verbs. Phrasal verbs have highly context-dependent meanings. Phrasal verbs produce specialized context-specific meanings that may not be derived from the meaning of the constituents.

EBMT also can be used to determine the context of the sentence.

Hybrid MT

Hybrid machine translation (HMT) combines the strengths of different translation methodologies. The approaches can be used in a number of ways:

- Rules post-processed by statistics: Translation is performed using a rules based engine. Statistics are then used in an attempt to improve the output from the rules engine.
- Statistics guided by rules: Rules are used to pre-process data in an attempt to better guide the statistical engine. Rules are also used to post-process the statistical output to perform functions such as normalization.

Approaches to evaluating the quality of machine translation

A measurement of the quality of the machine translation output is usually called a metric. The task for any such metric is to assign scores of quality in such a way that they correlate with human judgment of translation quality.

The measure of evaluation for metrics is correlation with human judgment. This is generally done at two levels, at the sentence level, where scores are calculated by the metric for a set of translated sentences, and then correlated against human judgment for the same sentences. And at the corpus level, where scores over the sentences are aggregated for both human judgments and metric judgments, and these aggregate scores are then correlated.

Good performance of a metric, across text types or domains, is important for the reusability of the metric. A metric that only works for text in a specific domain is useful, but less useful than one that works across many domains. Another important factor in the usefulness of an evaluation metric is to have good correlation, even when working with small amounts of data.

Attributes that a good automatic metric should have are:

- correlation;
- sensitivity;
- consistency;
- reliability;
- generality.

Nowadays there are several approaches used to evaluating the quality of machine translation.

BLEU

BLEU (Bilingual Evaluation Understudy) is an algorithm for evaluating the quality of text which has been machine-translated from one natural language to another. Quality is considered to be the correspondence between a machine's output and a human-translator's output. BLEU was one of the first metrics to achieve a high correlation with human judgments of quality and remains one of the most popular.

Scores are calculated for individual translated segments—sentences—by comparing them with a set of good quality reference translations. The scores are then averaged over the whole corpus to reach an estimate of the translation's overall quality. Intelligibility or grammatical correctness is not taken into account.

BLEU is designed to approximate human judgment at a corpus level, and does not show good results when used to evaluate the quality of individual sentences.

BLEU’s output is always a number between 0 and 1. This value indicates how similar the candidate and reference texts are, with values closer to 1 representing more similar texts.

BLEU uses a modified form of precision to compare a candidate translation against multiple reference translations. The metric modifies simple precision:

$$P=m/w_t$$

where m is number of words from the candidate that are found in the reference, and w_t is the total number of words in the candidate.

In order to produce a score for the whole corpus the modified precision scores for the segments are combined, using the geometric mean multiplied by a brevity penalty to prevent very short candidates from receiving too high a score.

BLEU has frequently been reported as correlating well with human judgement, and remains a benchmark for the assessment of any new evaluation metric. There are however a number of criticisms that have been voiced. It has been noted that although in principle capable of evaluating translations of any language, BLEU cannot in its present form deal with languages lacking word boundaries.

It has been argued that although BLEU has significant advantages, there is no guarantee that an increase in BLEU score is an indicator of improved translation quality.

NIST

NIST is a method for evaluating the quality of text which has been translated using machine translation. Its name comes from the US National Institute of Standards and Technology.

It is based on the BLEU metric, but with some alterations. Where BLEU simply calculates n-gram precision adding equal weight to each one, NIST also calculates how informative a particular n-gram is. That is to say when a correct n-gram is found, the rarer that n-gram is, the more weight it will be given.

NIST also differs from BLEU in its calculation of the brevity penalty insofar as small variations in translation length do not impact the overall score as much.

Word error rate

Word error rate (WER) is a common metric of the performance of a speech recognition or machine translation system.

The general difficulty of measuring performance lies in the fact that the recognized word sequence can have a different length from the reference word sequence. The WER is derived from the Levenshtein distance, working at the word level instead of the phoneme level.

This problem is solved by first aligning the recognized word sequence with the reference word sequence using dynamic string alignment.

Word error rate can then be computed as:

$$WER = \frac{S + D + I}{N}$$

where
S - number of substitutions,
D - number of the deletions,
I - number of the insertions,
N - number of words in the reference.

When reporting the performance of a speech recognition system, sometimes word accuracy (WAcc) is used instead:

$$WAcc = 1 - WER = \frac{N - S - D - I}{N} = \frac{H - I}{N}$$

where

H is N-(S+D), the number of correctly recognized words.

METEOR

METEOR (Metric for Evaluation of Translation with Explicit ORdering) is a metric for the evaluation of machine translation output. The metric is based on the harmonic mean of unigram precision and recall, with recall weighted higher than precision. It also has several features that are not found in other metrics, such as stemming and synonymy matching, along with the standard exact word matching. The metric was designed to fix some of the problems found in the more popular BLEU metric, and also produce good correlation with human judgement at the sentence or segment level This differs from the BLEU metric in that BLEU seeks correlation at the corpus level.

As with BLEU, the basic unit of evaluation is the sentence, the algorithm first creates an alignment between two sentences, the candidate translation string, and the reference translation string. The alignment is a set of mappings between unigrams. A mapping can be thought of as a line between a unigram in one string, and a unigram in another string. The constraints are as follows; every unigram in the candidate translation must map to zero or one unigram in the reference translation and vice versa. In any alignment, a unigram in one string cannot map to more than one unigram in another string.

Each stage is split up into two phases. In the first phase, all possible unigram mappings are collected for the module being used in this stage. In the second phase, the largest subset of these mappings is selected to produce an alignment as defined above. If there are two alignments with the same number of mappings, the alignment is chosen with the fewest crosses, that is, with fewer intersections of two mappings. From the two alignments shown, alignment (a) would be selected at this point. Stages are run consecutively and each stage only adds to the alignment those unigrams which have not been matched in previous stages. Once the final alignment is computed, the score is computed as follows: Unigram precision P is calculated as:

$$P = \frac{m}{w_t}$$

Where m is the number of unigrams in the candidate translation that are also found in the reference translation, and wt is the number of unigrams in the candidate translation. Unigram recall R is computed as:

$$R = \frac{m}{w_r}$$

Where m is as above, and wr is the number of unigrams in the reference translation. Precision and recall are combined using the harmonic mean in the following fashion, with recall weighted 9 times more than precision:

$$F_{mean} = \frac{10PR}{R + 9P}$$

The measures that have been introduced so far only account for congruity with respect to single words but not with respect to larger segments that appear in both the reference and the candidate sentence. In order to take these into account, longer n-gram matches are used to compute a penalty p for the alignment. The more mappings there are that are not adjacent in the reference and the candidate sentence, the higher the penalty will be.

In order to compute this penalty, unigrams are grouped into the fewest possible chunks, where a chunk is defined as a set of unigrams that are adjacent in the hypothesis and in the reference. The longer the adjacent mappings between the candidate and the reference, the fewer chunks there are. A translation that is identical to the reference will give just one chunk. The penalty p is computed as follows,

$$p = 0.5 \left(\frac{c}{u_m} \right)^3$$

Where c is the number of chunks, and um is the number of unigrams that have been mapped. The final score for a segment is calculated as M below. The penalty has the effect of reducing the Fmean by up to 50% if there are no bigram or longer matches.

$$M = Fmean(1 - p)$$

To calculate a score over a whole corpus, or collection of segments, the aggregate values for P, R and p are taken and then combined using the same formula. The algorithm also works for comparing a candidate translation against more than one reference translations. In this case the algorithm compares the candidate against each of the references and selects the highest score.

2 Techniques of semantic modeling

In the machine translation system based on rules, the source code the first thing analyzed morphologically and syntactically, to get a syntactic representation. Various methods of analysis and transformation can be used to obtain the final result. To choose methods and emphases are heavily dependent on system design, however, most systems include at least the following steps: morphological analysis, lexical categorization, lexical conversion, structural transformation, morphological synthesis, syntactic transformation (surface), semantic transformation (deep). Dwell in particular on semantic change and semantic modeling techniques. First language semantics - is a branch of linguistics which studies the semantic aspect of language that is meaning, the meaning of linguistic units (morphemes, words, phrases, etc.)

Semantic transformation (deep). This creates a level of semantic representation, which depends on the source language. This presentation may consist of a series of structures that represent meaning. In these systems, the translation is usually done predicates. Translation is also usually requires a structural transformation. This level is used to translate between more distant languages, or languages that have had no genetic relationship (Spanish - English or Spanish - the language of the Basques, etc.).

Methods of formal semantics

There are many methods, but the most famous and popular are the following:

Method of component analysis

In English, a separate line of compositional semantics. It was assumed that by a finite set of semantic components can be described as an unlimited set of lexical items. Technique selection semantic factors is to consider the allocation of certain words and signs, dividing the words into different classes and semantic groups, for example, on grounds such as animate / inanimate, male / female gender, etc. can be identified and more differentiated features for word classes, such as animals, fish, birds, people, etc. Meaning of each word thus appears as the set of semantic factors.

Consider a concrete example of this method. Take, for analysis of the word "journal". First, we must find a word or phrase indicating the kind of things, which is a kind of journal. This word - a periodical.

1. The value of this generic terms (hyperons) is the first semantic component within the definition of the word "journal". It displays the general features of the magazine with other publications of this kind (2 trait = frequency of publication). These common symptoms are called integral semantic features.

2. Search for all words denoting other kinds of periodicals and identify those attributes for which logs are different from other kinds of periodicals. Such signs are called differential semantic features.

In addition to magazines, periodicals are newspapers, newsletters, catalogs. From newspapers magazines differ in that they are stitched. If printed publications are not stitched, it is not a magazine. From the newsletter and directory journal is different on other grounds not related to registration of the publication and its contents. For example, create directories for publishing data about the product. Thus, the interpretation of the word "journal" includes, besides the integral sign, 2 differential. For the magazine are the components that characterize it from the look and feel of the content.

Method of component analysis is actively developing overseas. There are different theories of compositional semantics (Katz and Fodor scientists - pioneers, Barbara Patty, Anna Vezhbitskaya, etc.). For example, Vetbitskaya comes from the fact that the values of all words in all languages can be described using the same limited set of elements as irreducible atoms in physics, ie semantic primitives: many pronouns, numerals, verbs (doing, being able to think, speak, to know, have), size (large and small), adverbs (where, when), etc.

Basically, for the method of component analysis distinguished the thesis that "the meaning of sentences is the sum of meanings of its constituent words."

The method of semantic cases

Great contribution to the development of language for writing semantic structures and forms, introduced by Charles Fillmore. He accepted the hypothesis component structure values and the idea of sequential expansion of word meaning into simpler components up to the semantic primitives or atoms of meaning. Sharing the common views on the predicate argument structure, he concludes that it is necessary to specify not only the number of arguments of the predicate, but their role is semantic content. He identifies the following roles:

1. agent - animate the initiator of the action
2. object - a thing which is the subject of
3. counterparty - the force against which the action
4. recipient - the person for which an action
5. tool - the physical cause of action / motivation
6. source - the original state of the object to the action.

He also offered a detailed concept of lexical meaning. He is a classic of lexical semantics abroad.

The common conception of lexical meaning based on the concepts of layering, ie, includes shades of meaning, stylistically and emotionally expressive elements of values.

He goes further and adds value in two parts: the actual value and presubpozitsiya. For example, in saying "Vasya - not a bachelor" does not assert that Vasya was not a man. Ie if we assume that the word "bachelor" - an adult male never-married, then the negative is only the second part after the comma, which is the actual value.

The main result of this research is to review Fillmore usual scheme of entries in dictionaries. He believes the primary means of vocabulary task of semantic role structures and rules of their transfer to surface structures that are common to the concepts of such structures in the Russian research management models YD Apresian.

This theory of semantic cases and semantic dictionaries developed in the Moscow school of semantics, which created a model "text - the meaning - the text" and in particular the explanatory combinatorial dictionary of modern Russian language. For such a dictionary was developed apparatus of lexical features that are similar in the sense of the unit semantic cases Filmore. In 1984 was published version of the dictionary "New Explanatory Dictionary of Synonyms of the Russian language."

Semantic networks

Technology of semantic networks can lead civilization to a new level. It is for the reason that the government of U.S.A., France, Germany and other countries are investing in these developments a huge fiscal resources. These technologies are expected to address the context and as a result - the establishment of information systems, artificial intelligence. Programming will be possible in natural languages, the creation of smart weapons of the battlefield, advanced search and expert systems, and much more.

Unambiguous definition of the semantic web is not currently available. In knowledge engineering it’s means a graph showing the meaning of the complete image. Graph nodes correspond to concepts and objects, and arcs - relationship between objects. Formally, the network can be defined as follows: $H = < I, C, G>$

- **I** - a set of information units;
- **C** - many types of relationships between information units;

- **G** - the map that defines the specific relationship of existing types between elements. Semantic Web as a model most often used to represent declarative knowledge.

One of the first well-known models based on the semantic web is a TLC-model (Teachaple Languge Compre-hender - affordable mechanism for understanding the language), developed by Kuillian in 1968. Model was used to represent the semantic relationships between concepts (words) to describe the structure of long-term memory in human psychology.

Semantic networks, designed as a common unit of knowledge representation, from the very beginning actively used for the construction of systems of natural language processing. We consider various ways of representing cognitive content of statements in the NL with semantic networks. The simplest - the logical predicate: typed semantic relation connects two nodes. Since this representation does not allow to denote events in the node-set is turned on and the predicate itself and the circumstances of the action associated with it a set of relations. Semantic network of this type, shown in Figure 2.1, describes the importance of English phrases *On Wednesday morning John hit Mary in the park by the fist* as a set of nodes corresponding to objects or conceptual notions and related some of directed bonds.

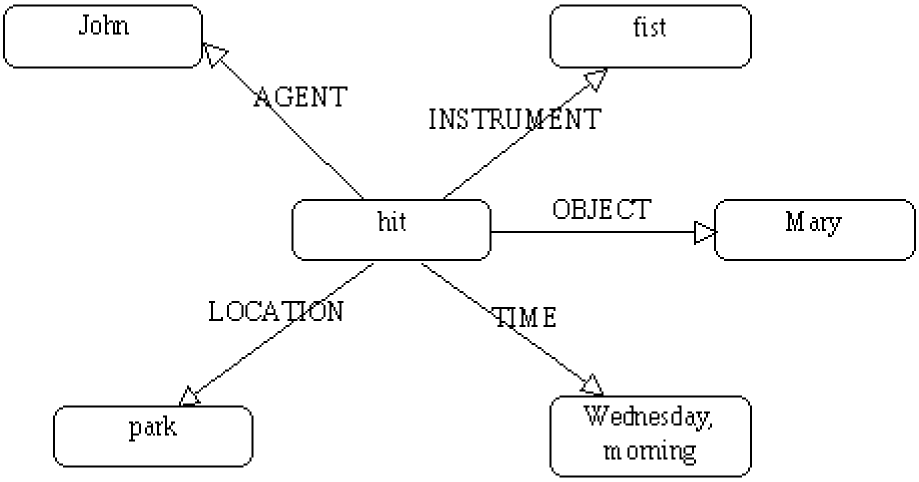

Figure 2.1

G.S. Tseitin offered his version of the presentation value of ER-expressions with the help of semantic networks. His associative networks contain nodes that express some entity (the relevant objects in the text or in the outside world), and directed arcs connecting these nodes. Site content may be a number, a string of symbols, the procedure or a finite set of other nodes. Arcs are named, and the names of all arcs leaving the node must be distinct, often they are intended to semantic and syntactic roles. In particular, the words "open" and "close" can be used to refer action with a wide range of objects: open / close door, a bottle, a computer file. In terms of associative networks, these words can also be represented as arcs that connect some types of objects and method for appropriate action.

System SNePS (Semantic Network Processing System) is widely used as a means of developing pilot applications using natural language, since it includes basic reporting mechanism and the automatic construction of semantic networks with minimal structure and mechanism of the withdrawal on those networks. The mechanism of ATN-grammars can also be described in terms of SNePS (ATN-grammar of English is part of the basic package SNePS). The close connection of this system with processing tasks ER is also due to the fact that architecture is not focused on the manual creation of semantic networks, and on its construction as a result of extracting knowledge from various sources, often ER-texts.

Theory of rhematic graphs usually use the mechanism of network representations to reflect the phenomena characteristic of language as a communication tool rather than as a mechanism for modeling the knowledge about the world (ontology). This theory is based on the mathematical theory of lattices (partially ordered sets) and provides a view of both the semantic, syntactic and phonetic information. Two rhematic graph, the relevant proposals "Larry was reading some trash" and "Larry was reading a comic bought at the station" and depicted in Figure 2.2, constructed from the parse tree of the proposals by converting it into an acyclic graph. This process involves the fusion of the leaf nodes belonging to the same object, and making order in accordance with the focus of attention: in Figure 2.2, dashed lines link the reference point (point of reference) and a point of interest.

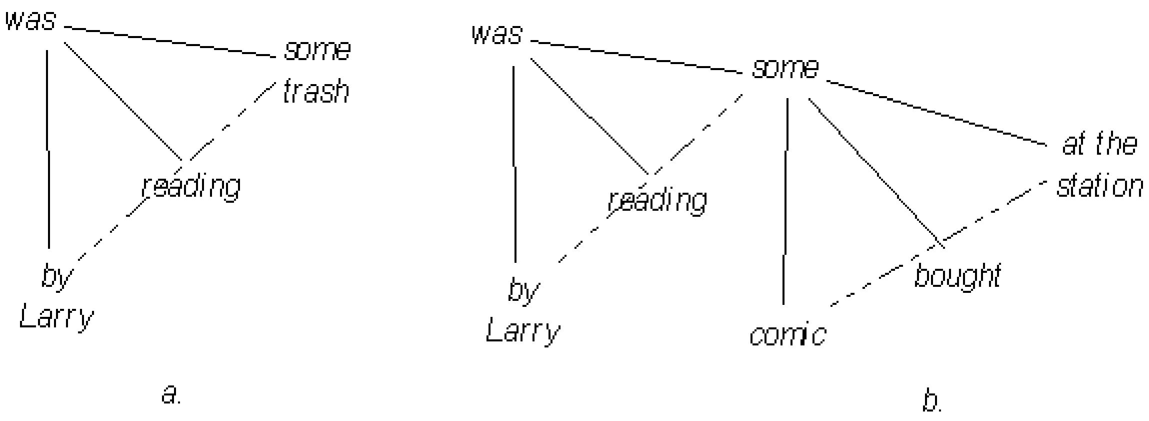

Figure 2.2

Based rhematic graphs combined into a single view when analyzing a text (see the analysis of the proposal "Larry was reading a comic he had bought at the station" in Figure 2.3), is a logical conclusion by modus (passive-active relation between action and situation), the type of operation (primary, secondary, tertiary) and arrays containing the classical class-subclass relations and phase (direction of action, including source, purpose and position).

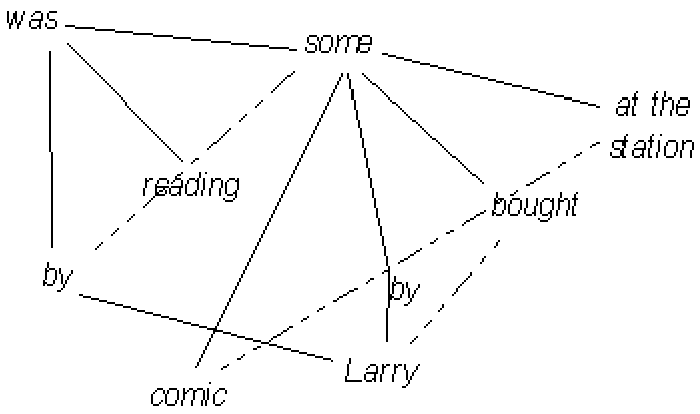

Figure 2.3

Joint lattice of these relationships is presented in Figure 2.4.

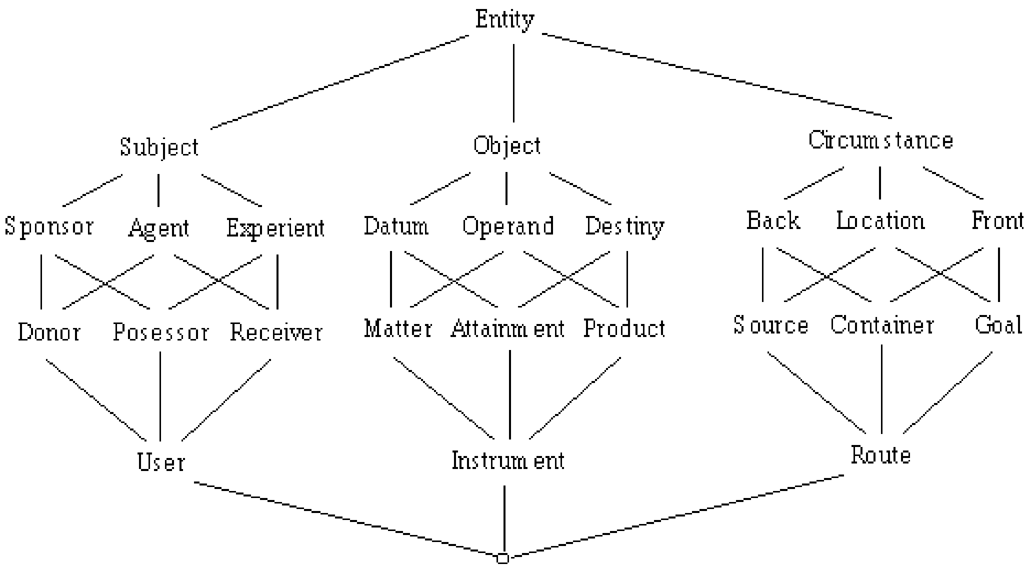

Figure 2.4

System SNOOP (System with Networks and Object-Oriented Productions), offers another way to represent semantic networks for language problems by integrating network representation with the methods of object-oriented programming and production rules. The network nodes and relations between them belong to the classes, described by a separate network inheritance in which the class is defined by a set of possible fields (which allows nodes to have an internal structure) and properties associated with objects relevant to this class. The properties of objects of classes described by the groups of production rules, each of which consists of the sample, allowing to navigate through network conditions at the sites found by the sample, and actions to change the internal structure of these sites and change the network (creation and deletion of nodes and relations).

Classification of semantic networks

For all semantic networks is true separation of the arity and the number of types of relationships.

• The number of types of relationships, the network can be **homogeneous** and **heterogeneous**.

-Homogeneous networks have only one type of relationship (arrows), for example, such is the above-mentioned classification of species (with the only attitude).

-In heterogeneous networks the number of types of relationships more than two. The classic illustration of this model of knowledge representation represent just such a network. Heterogeneous networks are of great interest for practical purposes, but also a great challenge for research. Heterogeneous networks can be represented as a tree-like interweaving of multilayer structures. An example of such a network may be Semantic Web Wikipedia.

Number of types of relationships in the semantic network is defined by its creator, based on specific goals. In the real world, their number tends to infinity. Each relationship is, in fact, a predicate, simple or compound. Speed of work with the knowledge base depends on how effectively implemented treatment programs necessary relations.

• The arity:

- are typical network with **binary** relationships (connecting exactly two concepts). Binary relations are very simple and convenient to represent the graph as an arrow between two concepts.

In addition, they play a crucial role in mathematics.

- In practice, however, may need relationships that connect more than two objects - **N-ary**. This gives rise to complexity - how to portray such a relationship on a graph, not to be confused. Conceptual graphs remove this difficulty by presenting each relation in the form of a single node.

In addition to conceptual graphs, there are other versions of semantic networks, this is another basis for classification (**implementation**).

Features of software development for natural language processing

Based on what was described above it is possible to make some conclusions about features that natural language processing software definitely needs to have. At first glance a theory of the approaches seems to be enough to produce good software tools for the tasks of NLP. But as the state of the art of such software tools shows, it is not as easy. Development of such software requires some complex features that are described bellow.

Development of machine translation software

- Software for MT should be developed on a modular basis. As translation process usually consists of several stages.
- Texts to be translated should be analyzed not only on sentence level, but on phrase (collocations or n-grams) and word level as well.
- When using rule-based approaches some detailed information about the grammar rules of source and target languages is very important. This is why MT software will differ for different language pairs. And translation of every pair should be viewed separately, as different programming task.
- Statistical and example based approaches require linguistic corpora available for every language involved and bilingual corpora for language pairs. The larger corpora are, the better the quality of translation is. If there is no corpora or very few of them for language(s) involved in translation, then probably the task of creating some of such structures should be considered. Which is not an easy task itself.
- Hybrid approach, though it may seem the optimal one, is quite dangerous because without proper linguistic modeling it may combine disadvantages of approaches used.

As an example of machine translation software developing we can provide a Kazakh-English machine translation system. The system is developing in Kazakh National University since 2008. It has following characteristics:

- Machine translation from Kazakh into English is done with studying linguistic features of both languages.
- Rule-based approach and Statistical approach to machine translation cannot produce good translation separately, that is why hybrid approach is being used.
- In order to produce translation as close to publication quality as possible MT system has pre-editing of source text.
- As there is no corpora for Kazakh language MT system will have some tools to help to produce monolingual and bilingual corpora.
- MT system has following parts (modules):
  - sentence boundary identification;
  - length check;
  - morphological analyzer;
  - syntactical parser;
  - bilingual dictionary;
  - morphological generator;

    - syntactical generator;
- Morphological analyzer:
    - splits sentences into words;
    - carries out vocabulary control;
    - analyzes words for morphological features;
    - uses regular grammar for parsing;
    - saves unknown words for future adding to a dictionary;
- Syntactical parser
    - performs part-of-speech tagging;
    - outputs scheme of the sentence in source language;
    - uses Link grammar theory;
    - parsing success control;
- Bilingual dictionary:
    - consists of several dictionaries (general, term-base, named entities);
    - contains word pairs and collocation pairs in Kazakh and English;
    - contains morphological and syntactical data that cannot be described in grammars;
- Morphological generator:
    - generates words in corresponding morphological form in target language;
    - uses regular grammar to generate word forms;
    - uses a statistical module for word sense disambiguation;
- Syntactical generator:
    - produces scheme of the sentence in target language;
    - puts translated words and collocations in target language into the scheme;
- Source text and output target text are saved for human revision and future adding to monolingual and bilingual corpora.

Morphological analyzer uses regular grammar for parsing separate words. Which is quite usual usage of that type of grammars. Syntactical parser instead of context-sensitive grammars uses Link Grammar theory for parsing sentences.

Link grammar is a theory of syntax which builds relations between pairs of words, rather than constructing constituents in a tree-like hierarchy. There are two basic parameters in the link grammar: directionality and distance. Here is an example of how this grammar is used in parsing.

Sentence: **Марат жаңа кино көрді.**

Simple link grammar rule describing this type of sentences looks like this:

<subject>: **S+;**
<adjective>: **A+;**
<object>: **A- & O+;**
<verb>: **S- & O-;**

It means that in this type of sentences <subject> can have 1 link of S type from the right, <object> can have 1 link of A type from the left an 1 link of O type from the right and so on.

Parse tree of the sentence:

```
+-------S-------+
|    +--A-+--O-+
|    |    |    |
```

Марат жаңа кино көрді.

Corresponding link grammar rule for English translation, that will be produced by syntactical generator:

```
<subject>:          S+;
<verb>:             S- & O+;
<determiner>:       D+;
<adjective>:        A+;
<object>:           D- & A+;
```

Translation: **Marat watched a new movie.**

Parse tree for sentence in English:

```
       +------O-----+
       |    +---D---+
  +---S--+   | +--A-+
  |      |   | |  |
Marat watched a new movie.
```

This simple example confirms that machine translation is not as trivial task as it seems from the first look.

Development of quality evaluation software for machine translation

Quality evaluation software is not as difficult as MT software. But still it has its own characteristics. Bilingual corpora specific for domain used in translation are needed for program learning and performing the evaluation. Also, because of the algorithms' features described above, supervision of a professional translator when conducting an evaluation is highly recommended.

References

Daniel Jurafsky, James H. Martin *SPEECH and LANGUAGE PROCESSING* Prentice Hall, 2008

Adam Boretz, *AppTek Launches Hybrid Machine Translation Software* SpeechTechMag.com, 2009

Bogdan Babych, Anthony Hartley, and Serge Sharoff *Translating from under-resourced languages: comparing direct transfer against pivot translation.* Proceedings of MT Summit XI, 10–14 September 2007

Papineni, K., Roukos, S., Ward, T., and Zhu, W. J. *BLEU: a method for automatic evaluation of machine translation* ACL-2002: 40th Annual meeting of the Association for Computational Linguistics

Banerjee, S. and Lavie, A. *METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments* Proceedings of Workshop on Intrinsic and Extrinsic Evaluation Measures for MT and/or Summarization at the 43rd Annual Meeting of the Association of Computational Linguistics, 2005

Daniel Sleator, Davy Temperly *Parsing English with a Link Grammar* Third International Workshop on Parsing Technologies, 1993

Authors names and affiliations

Ualsher Tukeyev
al-Farabi Kazakh National University
Almaty, Kazakhstan
Ualsher.Tukeyev@kaznu.kz

Zhandos Zhumanov
al-Farabi Kazakh National University
Almaty, Kazakhstan
z.zhake@gmail.com

Diana Rakhimova
al-Farabi Kazakh National University
Almaty, Kazakhstan
di.diva@mail.ru