Meaning Understanding in Machine Translation

Hirosato Nomura

Department of Artificial Intelligence
Kyushu Institute of Technology
Iizuka-shi, 820, Japan

Abstract

We have been studying the meaning-based approach to machine translation which can account for the generalization of the idea based on the thinking way of relationships between two objects in a language expression and/or deep structures embedded in it. In our approach, this came from the idea of the case relation in Case Grammar while the similar idea might have been adopted to design many things in various area. This idea is very fundamental in designing a linguistic and computational model for machine translation or more generally natural language processing including Japanese language processing, and can be extended naturally to the relationships between more complicated and various kinds of objects such as concepts, lexical items, phrasal components, clausal components, contextual components, and even knowledge fragments and discourse and/or situations. Thus, this paper tries to describe a computational linguistic model based on the idea of the relationships which will be adapted to describing lexical, grammatical, semantic and contextual information in a language expression.

In the beginning, this paper describes the outline of the memory structure and its representative framework on which the meaning understanding is carried out. This memory structure constructs the meaning structure of a text step by step by analyzing each sentence in the text and then assimilating the result into the memory structure. Secondly, it defines syntactic, semantic and contextual relationships in a language expression which combine meaning structures corresponding to the components in the language expression and then construct a larger meaning structure. This is done by extending the idea of the case relation between a verb and a case element to that on various kinds of objects varying from lexical objects and contextual objects. Next, it describes other clues which are needed for completing the definition of the meaning structure. They include grammatical and semantic primitives such as syntactic and semantic categories. Then, it defines the meaning structure ultimately which represents the computational linguistic model for meaning understanding we propose in this paper. Finally, it describes the application of the model to the design of the analysis of a text for machine translation.

1. Introduction

There are two approaches in developing machine translation technologies: One is to assemble current technologies related to computational linguistics, natural language processing, or simply machine translation, by which we can produce a practical machine translation system even if it is realized by the insufficient technologies. The other is to try to study expectable technologies for future by which we can get the technical prospects for a high ability and/or quality machine translation system. This paper is intended to describe one of the second approaches and concerns mainly a framework for meaning understanding of a text which will be adopted to design a meaning-based machine translation system.

There have been conducted many interesting machine translation projects over the world during last twenty-five years such as ARIANE and CALLIOPE in France, SUSY and SEMSYN in West Germany, TAUM in Canada, Mu and others in Japan, EUROTRA in CEC, ALVEY's in United Kingdom, CMU's and XTRA in USA, ROSETTA and DLT in Holland, and so on. Some of them concern the traditional syntactic approaches while some of the others concern the adaptation of the newly developed linguistics theories, and some of the rest concern the so-called knowledge-based or similar approaches. Among such distinguished projects listed above, many seem to have been approaching from both first and second standpoints while some of them stress producing a practical machine translation system and some others stress developing new technologies.

We have been studying the meaning-based approach to machine translation which can account for the generalization of the idea based on the thinking way of the relationships between two objects in a language expression and/or deep structure embedded in it. In our approach, this came from the idea of the case relation in Case Grammar while the similar idea might have been adopted to design many things in various area. Thus, this idea is traditional in its nature, however, this is very fundamental in designing a linguistic and computational model for machine translation or more generally natural language processing and can be extended naturally to the relationships between various kinds of objects such as concepts, lexical items, phrasal components, clausal components, contextual components, and even knowledge fragments and discourse or situations. Also it can be extended to the relationships between corresponding objects in different languages, which is essential for machine translation. Moreover, the idea of the so-called grammatical function widely appeared in the recent interesting linguistics theories also relates to this idea while it is rather active in a sense that it finds or bridges to a partner by which they can establish a relationship between an object and its partner as a result. Also the conceptual dependencies can be seen as an extension of the idea in a sense. However, the detail considerations of such extensions are beyond the scope of this paper.

This paper tries to describe a computational linguistic model for meaning understanding of a text based on the idea of the relationships, which will be adapted to design a framework for the uniform representation of lexical, grammatical, semantic and contextual information in a language expression. It describes, in the beginning, the memory model and its representative framework. Secondly, it defines syntactic, semantic, and contextual relationships. Next, it gives a linguistic model by applying the relationships. And, finally, it describes the application of the model to the design of the analysis of syntactic, semantic and contextual structure of a text. This paper concerns only the reference and ellipsis as the contextual problem and

leave further ambitious problems such as more general contextual analysis, discourse analysis, and situation analysis.

2. Understanding Model

2.1 Memory Structure

It is convenient and natural for our approach to define a memory structure for meaning understanding by three hierarchical structures: long-term, medium-term and short-term memory whose roles are explained shortly below.

The long-term memory is a static knowledge-base and stores both linguistic and non-linguistic knowledge. Linguistic knowledge relates to lexicon and grammar and non-linguistic knowledge relates to common sense and expertise. Selectional constraints or semantic conditions to be used for grammatical analysis are also involved in the knowledge-base while some part of it belongs to linguistic knowledge and the rest part belongs to non-linguistic knowledge. The distinction between linguistic and non-linguistic is not so significant for meaning understanding, however, it is needed for defining an isolated grammatical model.

The medium-term memory is a dynamic knowledge-base instantiated from the long-term memory in an environment and stores a speaker and a hearer model in an utterance situation. The medium-term memory acts as a discourse memory or a script-like memory and incorporates text meaning including both intra-sentential and inter-sentential information. The intra-sentential information accounts for the syntactic and semantic relationships holding among components in a sentence while the inter-sentential information accounts for the relationships holding among components belonging to different sentences in a text or different clauses in a complex sentence.

The ellipsis and reference are inter-sentential in nature and those relationships are analyzed through assimilation of the short-term memory into the medium-term memory. The analysis of their relationships is a part of the contextual analysis, thus the contextual analysis is accomplished by the interaction between the short-term memory and the discourse memory. This process thus completes the meaning understanding of a text which we concern here since we exemplify only the ellipsis and reference as examples of contextual analysis and we do not concern the further problems such as those related to discourse and/or script analysis. After completing the meaning understanding process, newly acquired information is assimilated back into the long-term memory if it is intended to do so.

The short-term memory contains instantaneous data obtained from the analysis of each sentence or each phrase in a sentence and temporarily stores the syntactic and semantic relationships among components of an ongoing sentence apart from the context and discourse or situation. This memory structure thus does not concern inter-sentential information.

2.2 Representative Framework

The representative framework of the memory structure we adopted here is essentially a semantic network consisting of nodes and arcs while each node is generalized to a frame, thus, the semantic network is represented as a frame-network.

Each node or each frame represents information concerning a component of a text which might be a word, a phrase, a clause, a sentence, or a sequence of sentences.

A frame consists of several slots. One of them contains a sub-frame network indicating its internal construction while it might be null when the node represents a word or something like a primitive unit. Since the sub-frame network is also a frame network, a frame network has an embedded or recursive structure which usually reflects the syntactic structure of a component in a text. The rest of them store values of features assigned to the node. Thus, the set of them indicates the values of the feature bundles given to the node and each value is summed up by some operations such as unification from those of nodes involved in the sub-frame network.

3. Relationships

3.1 Syntactic Relationships

Syntactic relationships combine components of a text grammatically and then produce a bigger component. For Japanese language, as an example of the so-called non-configurational language, however, we do not assume any syntactic structure excepting modification or dependency structure which can be seen as a directed relationship or a directed arc in a frame network representation. For English, as an example of the so-called configurational language, we assume grammatical functions as seen in LFG and GPSG, etc., which can also be seen as a directed relationship. As a result, we assume the directed relationships for both type of languages, thus those directed relationships make a directed frame network or a directed (acyclic) graph like in PATR-II. However, we do not analyze the pure syntactic structure of the component, instead, we analyze the semantic structure which might be parallel with the syntactic structure, as argued in the literatures concerning CUG. Thus, any syntactic relationship accompanies a semantic relationship in parallel which might be thought as a selectional restriction or grammatical function. It is not important here whether some of the grammatical functions have to be classified into as syntactic or semantic.

3.2 Semantic Relationships

Semantic relationships combine semantic information represented in frames and then build a larger frame network. The semantic relationships are independent from contextual and situational restrictions here, however, they will be regulated by inter-sentential relationships to be specified in a bigger frame network.

Taxonomic relation is a semantic relationship between semantic features, and it provides a basis for the classification of the semantic features.

Noun relation is a semantic relationship between nouns and is exemplified as whole-part, upper-lower, possession, and material, etc.

Case relation is a semantic relationship between a case element and a predicate and can be exemplified as object, agent, instrument, and place, etc. They are the well known relations and used widely for case analysis.

Embedded relation is a semantic relationship between an embedded sentence and a noun phrase, which can be categorized into three types; a) case relation between a modified noun phrase and the predicate in a modifier embedded sentence, b) noun relation between a modified noun phrase and a noun phrase in a modifier embedded sentence, and c) an appositive or subsidiary relation between a modified noun phrase and a modifier embedded sentence.

Coordination relation is a relationship which combines two components both have the same syntactic role in the bigger component. This might be seen as a special relationship of conjunctive relation mentioned below.

Conjunctive relation is a relationship between clauses or sentences, and can be exemplified as cause-result, time-advance, and assumption, etc.

3.3 Contextual Relationships

As mentioned earlier, we concern here only the reference and ellipsis as contextual relationships which can be seen as the basic relationships of the cohesive relationships. Such cohesive relationships are inter-sentential and give additional constraints on the semantic restrictions. Thus they resolve the ambiguities which could never be removed by the intra-sentential constraints.

Reference is classified into coreference and indirect reference. Coreference is a relationship indicating the fact that linguistic expressions including pronouns and substitution expressions point to the same node in the underlying semantic network.

Indirect reference is a similar relationship, however, linguistic expressions do not point the same node directly but are related by a link established through inference on the frame network. Such inference is carried out by applying contextual information and non-linguistic knowledge including situational conditions.

Ellipsis is a relationship while one component to be related to is disappeared. Such a disappeared component is called a zero pronoun. Ellipsis occurs when a linguistic expression has already appeared in a preceding clause or sentence and its omission does not cause serious ambiguity in the present sentence. Such an ellipsis also makes coreference. Thus, the ellipsis is a coreference between a component and the disappeared component. In the case such that a verb is omitted, the word "zero pronoun" is not adequate, however, we do not concern it here.

4. Linguistic Model

4.1 Basis for Linguistic Model

As a basis for organizing linguistic model, we assume three kinds of primitives: structure, relation, and concept.

The structure represents an internal construction of a component of a text and is thus stored in the slot for the sub-frame network. The primitive structure is provided for a word and the compound structure is provided for a phrase, clause, sentence, or text. The relation is the syntactic, semantic, and contextual relationship as mentioned above and combines the structures to produce a bigger structure. The

concept is used for describing the values of a feature bundle given to a frame, and the set of values of the feature bundle given to the frame makes new concept for the whole frame.

4.2 Dictionary

Lexical item stores such information as the structure and the concept, thus represents a frame for a word. Also it holds word oriented additional information which is helpful for ordering ambiguities, for example.

Semantic category is provided for specifying word meanings. Those for nouns and adverbs are used as selectional constraints in semantic relationship analysis. Those for predicates are used to analyze modality.

Case frame is provided for specifying predicate word meanings, thus, it retains a set of case relations which specifies the role and meaning of the verb. Each case relation is regulated by the constraints imposed on it. There are three types of case frames: intrinsic one for each predicate word meaning, common one for several predicate word meanings, and optional one for outer case relations.

Noun relation frame is provided for specifying word meanings of a complex noun. For predicate-type nouns, case frames are also used for specifying the relationship since its syntactic and semantic role in language expressions is similar to that of the corresponding predicate.

Event relation frame is provided for specifying predicate word meanings. An example of the relation appeared in the event relation frame is a relationship between a verb in a main clause and a verb in a subordinate clause.

Heuristics is used for resolving ambiguities among semantic categories, semantic relations, and semantic structures by linguistic information such as preference over several semantic relations. This includes some conditions on the semantic relations to be assigned to the relationship.

4.3 Structure Pattern

A structure pattern is a package of sub-structures, relationships combining these structures, and concepts represented by the set of the values of the feature bundle given to the structure. The structure pattern is provided for each construction of typical components or constituents.

A structure pattern consists of three parts: 1) the condition for applying the pattern, 2) the procedure for analyzing the internal construction, and 3) a structure type generated by the successful application of this structure pattern. The first part describes whether the structure pattern can be applied to a sub-structure sequence or a set of sub-structures. The second part performs a semantic relation analysis of the sub-structure sequence that satisfies the above condition. The third part describes the structure type to be produced by the above procedure.

A structure pattern might be viewed as a context free grammar, where the condition part corresponds to the right hand side of the CFG rule, the structure type part corresponds to the left hand side of it, and the procedure part can be seen as a procedure to derive the left hand side from the right hand side.

4.4 Non-Linguistic Knowledge

Both common-sense knowledge and expert-knowledge are also described by using basic primitives such as concepts, relations and structures mentioned earlier. Thus, non-linguistic knowledge is represented by the same framework provided for representing linguistic knowledge. However, some dedicated relationships and additional information are used for describing this type of knowledge.

Concept relation is a relationship such as hyponymy, synonymy, antonymy, whole-part, and possession. Event-State relation is a relationship between two events or between an event and a state. The subsidiary situation between "smell" and "grill" is an example since the "smell" results from the "grill."

Meta-knowledge is additional information used for reasoning, such as traversing frame networks, and checking semantic and contextual consistency according to frame networks.

4.5 Relational Structural Model

The traditional Fillmorean type case structure model is a model for representing the structure of a unit sentence which consists mainly of relations between case elements and a verb. Thus, this is a kind of the structure pattern specified by case relations.

The Relational Structural Model (RSM) proposed here is a multi-fold extension of the traditional case structure. It is defined based on the idea of relationships extended to those which bridge over various kinds of objects appeared as components in a text, as pointed out earlier. However, the relationships constructing a structure represented by RSM is intended to act as active relationships so that they find partners for the given components. By this formalization, RSM becomes to be able to represent attributes of a component and simultaneously its syntactic and semantic and even contextual structures by the directed relationships.

Though RSM is defined by semantic relationships under the assumption that it is language independent, it can retain language dependent structural information behind it since its sub-frame network reflects its syntactic construction. This is very convenient for machine translation because RSM provides the so-called inter-lingua representative framework for a language expression while it also provides an extra room for noting syntactic differences among languages, even one belongs to configurational languages and the other belongs to non-configurational languages.

5. Application to Machine Translation

5.1 Intra-Sentential Processing

The intra-sentential analysis is to analyze a sentence and to construct a frame network for the sentence. On the way, each word frame, each phrase frame and each clause frame are generated temporarily and then assimilated as sub-frame networks into the under-constructing sub-frame network of the component of the sentence or the under-constructing frame network of the whole sentence.

The structure pattern is used for predicting a syntactic relationship between a pair of a modifier and a modificant in a constituent structure of a component or a syntactic construction of the component. Based on this prediction, an analysis procedure is invoked to analyze their syntactic and semantic relationships in detail. If this analysis succeeds, a relationship for the pair of the modifier and modificant is recognized and the pair is integrated into a new structure by the relationship. Thus, the analysis seems to be syntax or object driven analysis in a sense, however, it is really semantics or restriction oriented analysis since well-formedness of each syntactic constituent is checked by semantic constraints simultaneously and the most effort in the analysis is devoted to the semantic check. The structure pattern merely navigates the semantic analysis.

Ambiguities induced by lack of contextual or situational information are left for the inter-sentential analysis. Detail strategies of the intra-sentential analysis consisting of noun analysis, case analysis, and modality analysis, etc. have appeared in the literature.

5.2 Inter-Sentential Processing

The inter-sentential analysis is defined as an assimilation process of the short-term memory into the medium-term memory. Main tasks of the inter-sentential analysis is to resolve cohesive restrictions which must be satisfied by the newly added intra-sentential frame network. Essentially, the intra-sentential analysis produces all of the possible candidates, thus, the inter-sentential analysis is to select one among them which gives reasonable interpretation of the sentence in the given context already represented as the frame network for the sequence of sentences appeared in the former part of the text.

In analyzing cohesive conditions, the cohesive relationships semantically combine two components directly or indirectly through reasoning accomplished by traversing generalized nodes over the frame network. Each of the cohesive relationship analysis is explained in the following three sections.

5.3 Analysis of Coreference

Analysis of coreference is to find two frames pointing a same object. However, even if two components are the same in the linguistic expressions in a text, they do not necessarily make coreference. For example, referred objects might be the same in class but different in type; one might be a prototype while the other might be an instance. Therefore, the type identification of the nouns, for example, is crucial to analyzing coreference between nouns and can be resolved by applying information on so-called topic and sentence types such as propositions or facts distinguished by time adverbials, tense information, and predicate meanings, while the detail discussions of these problems are beyond the scope of this paper since they relate deeply to the discourse information.

5.4 Analysis of Indirect Reference

Analysis of indirect reference is carried out by applying relationships successively to find a path from one frame to another frame on the frame network. However, such a process usually diverge, thus, careful selection of the useful relationships is needed for successful reasoning. Unfortunately, the strategy for it is

not clear, thus, all of the available conditions have to be considered simultaneously in the selection, and effective relationships for the objects have to be applied in the inference.

One example of such relationships is concept relation. Concept relation is a relationship between concepts represented by noun phrases, and is thus used in inference to determine indirect reference between the noun phrases. For example, the relation between "a *station name"* and *"the number of letters"* is inferred from the two concept relations: one denotes that "a *station name"* is an instance of *"name"* and the other indicates that *"name"* has the *"the number of letters"* attribute. Thus, the indirect reference between the noun phrases in this example can be easily resolved. However, in an example, it is rather difficult to analyze indirect reference between "a *transportation"* and *"ten letters"* since several concept relations must be applied to resolve the relationship between them, which find an exact link as a sequence of relationships between two frames in the hierarchical frame network.

Another example of the relationship which sometimes applied for resolving the indirect reference is event relation. Event relation is a relationship between events represented by sentences, and is thus used in inference to determine indirect reference between sentences. Event relations account for cause and result, motivation, indispensable prerequisite, examples, and elaboration, etc.

5.5 Analysis of Ellipsis

There are some types of ellipsis depending on what kinds of components are omitted in the language expression.

An elliptic obligatory case element in a sentence can often be identified semantically with a case element appearing in the preceding sentence in a sequence of sentences. This kind of elliptic element is exactly called a "zero pronoun". A zero pronoun and the preceding case element point to the same node on the semantic network, and thus make a coreference. For identifying an elliptic case element, syntactic and semantic constraints imposed on the unfilled case slot in a case frame, which is a slot for the omitted case element, can be applied. Additionally, so-called a topic, a focus, a point of a view, predicate meanings, and pragmatic constraints will also be applied effectively, however, the discussions of these conditions are again beyond the intended scope of this paper.

As for identifying an elliptic case element within a complex sentence, the characteristics of conjunctions or conjunctive particles can be applied effectively. Ellipsis of constituents indicating a speaker or a hearer occurs frequently and naturally in Japanese, and can be seen as a special case of zero pronouns. Such ellipsis can be resolved by using information such as honorific expressions, causative auxiliaries, and request expressions, while the detail discussions are omitted here.

An expression having a predicate is also omitted sometimes in a clause in a coordinating structure, and thus, induces another type of ellipsis. This elliptic expression can be related to a sub-frame network constructed by the expression appearing in the first clause. However, the frame as a case frame or a predicate type frame will never be constructed for the second clause without recognizing the fact that the second clause is really a clause and it involves an elliptic expression concerning the predicate. If it is recognized fortunately, a strategy can be adopted since the frame for the elliptic predicate must satisfy constraints imposed by conditions specified in

the filled case slots of the case frame assigned to the elliptic predicate in the second sentence.

A modifier phrase in a complex noun phrase may be omitted and only the head-phrase expressed, thus gives the next type of ellipsis. In such a case, functionality or grammatical function of the head noun is used for finding a referenced modifier phrase. Examples of the semantic relationships providing such functionality are whole-part, object-attribute, event-reason, and event-goal relationships.

References

[1] J. Bresnan: "The Mental Representation of Grammatical Relations", MIT Press, 1983.

[2] G. Gazdar, E. Klein, G. K. Pullum and I. A. Sag: "Generalized Phrase Structure Grammar", Basil Blackwell, 1985.

[3] W. J. Hutchins: "Machine Translation", Ellis Horwood Limited, 1986.

[4] H. Iida, K. Ogura, and H. Nomura: "A Case Analysis Method Cooperating with ATNG and its Application to Machine Translation", Proceedings of the International Conference on Computational Linguistics, 1985.

[5] S. Naito, A. Shimazu, and H. Nomura: "Classification of Modality Function and its Application to Japanese Language Analysis", Proceedings of the 23th ACL Annual Meeting, 1985.

[6] H. Nomura, S. Naito, Y. Katagiri, and A. Shimazu: "Translation by Understanding: A Machine Translation System LUTE", Proceedings of the International Conference on Computational Linguistics, 1986.

[7] A. Shimazu, S. Naito, and H. Nomura: "Japanese Language Semantic Analyzer Based on an Extended Case Frame Model", Proceedings of the International Joint Conference on Artificial Intelligence, 1983.

[8] A. Shimazu, S. Naito, and H. Nomura: "Semantic Structure Analysis of Japanese Noun Phrases with Adnominal Particles", Proceedings of the 25th ACL Annual Meeting, 1987.

[9] P. Sells: "Lectures on Contemporary Syntactic theories: An Introduction to Government-Binding Theory, Generalized Phrase Structure Grammar and Lexical-Functional Grammar", CSLI Lecture Notes, 3,1985.

[10] S. M. Shieber: "Separating Linguistic Analyses from Linguistic Theories", SRI International Tech. Note, 422,1987.

[11] S. M. Shieber: "An Introduction to Unification-Based Approaches to Grammar", CSLI Lecture Notes, 4,1986.

[12] M. Steedman: "Combinatory Grammars and Prastic Gaps, Edinburgh Working Papers in Cognitive Science Vol. 1: Categorial Grammar, Unification Grammar and Parsing", University of Edinburgh, 1987.