[*MT Summit III*, Washington, DC, USA, July 1991]

Multi-lingual Sentence Generation from the PIVOT Interlingua

Akitoshi OKUMURA Kazunori MURAKI Susumu AKAMINE

C & C Information Technology Research Laboratories
4-1-1 Miyazaki, Miyamae-ku, Kawasaki 216, JAPAN
{okumura,k-muraki,akamine}%mtl.cl.nec.co.jp@sj.nec.com

Abstract

This paper proposes a strategy for French and Spanish sentence generation systems, based on the English generation system. The English generation mode! consists of four procedures, conceptual wording (sentence-structure planning), syntactic selection, ordering and morphological generation. The analysis of linguistic similarities and differences between English, French and Spanish reveals that a single model is applicable for Spanish and French generation by modifying the contents of the dictionary and the knowledge base. The differences can naturally be represented in the knowledge base and dictionary, which is difficult in the transfer approach. So, it is possible to develop Spanish and French generation systems from the English generation system, merely by modifying the contents of the dictionary and knowledge base. Using this strategy, the authors have succeeded in developing English, Spanish, French, Korean and Japanese sentence generation systems. The input for each system is a language-independent conceptual representation, which is an interlingua of the machine translation system PIVOT. The sizes for each dictionary are 70,000 English words, 5,000 Spanish words, 3,000 French words, 110,000 Japanese words and 10,000 Korean words. The systems are implemented on NEC's work station.

1 Introduction

Multi-lingual machine translation system (MMTS) has been gaining more and more attention in recent years, There are two different methods for use in the system, a transfer approach and an interlingua approach. The interlingua method is suitable for MMTS, because of the low cost of the system development and maintenance[l]. From the viewpoint of system development effectiveness, the rules and dictionary information for the established system should be exploited for another language system development. When a sentence generation system is developed in one language, it is desirable to make the most use of the existing system, in order to develop another generation or analysis system more efficiently. For its purpose, linguistic phenomena of the languages should be analyzed from the semantic, syntactic and morphological view. It. is necessary to reveal their similarities and differences, A generation model should be designed to maximize the universal characteristics, but still to minimize language-specific characteristics.

The authors have comparatively analyzed subjects, sentence structures, grammatical cases, voice and morphological features in English, Spanish and French, and found it possible to develop Spanish and French generation systems from English, merely by modifying the contents of the dictionary and knowledge base. The generation model consists of four procedures; conceptual wording (sentence-structure planning), syntactic selection, word ordering and morphological generation. In each procedure, linguistic differences can be treated by the change in lexical information in the dictionary and knowledge bases. Most of the generation rules can be shared in the three languages.

This paper describes a strategy for developing the Spanish and French generation systems from one original English base system, according to the three languages comparative analysis. The input for the systems is a language-independent conceptual representation which the machine translation system PIVOT adopts as an interlingua[2, 3]. This strategy is applicable to Japanese and Korean generation[4]. This is reported in another paper. The authors succeeded in developing Spanish and French generation systems from an English generation system, and in developing Korean from Japanese. This paper first briefly explains an input interlingua. Next, a comparative analysis is shown for subjects and sentence structures in English, French and Spanish, and how to treat the differences. Then, the proposed multi-lingual sentence generation model is described.

2 PIVOT Interlingua

The PIVOT interlingua is a conceptual representation independent from language syntactic structure, which is an acyclic and directional network. All nodes are associated with conceptual primitives (CPs). It contains the following information.

1. Structural Information

All arcs have a direction. The direction of the arc shows semantic dependency relations.

2. Node Information
   All nodes have several kinds of semantic information, represented by a CP semantic symbol. This information can include such concepts as location, time, aspect, intention, objects, things, relations, etc. The authors call the relation CP (Agent, Object, Experiencer, etc) an involved semantic case, "a deep case".
3. Pragmatic Information
   Topic, Focus, Theme, Position as the argument of a predicate
4. Scope Information
   Scope of comparison, negation and quantification
5. CP-Structure Paraphrasing Postulates
   These are included as dictionary content and semantic postulates.

The authors developed English-Japanese and Japanese-English translation system PIVOT, which adopts the above mentioned interlingua method. The authors have established a multi-lingual sentence generation system, based on this interlingua method[3].

3 Lexical information for Spanish and French

Subjects, sentence structures, grammatical cases, voice and morphological features are comparatively analyzed in English, Spanish and French from the semantic, syntactic and morphological viewpoints. They are incorporated into the dictionary and some knowledge bases as lexical information. This section indicates how to describe them as lexical information about subjects and sentence structures,

3.1 Sentence structure and components

Components of a simple sentence and their basic order are shown in Table 1, where the component means a grammatical case.

Since predicate verbs carry obligatory grammatical cases as their arguments, they could be classified by arguments, structures of the arguments, permissible shifts and the relation among the arguments. According to Hornby's Verb Pattern and the authors' analysis, 111 verb patterns (VPs) are defined in English [5, 9]. Many English verb patterns are found to be available in Spanish and French.

An English predicate verb can take six grammatical cases (S, IOB, DOB, PP, COMP, CONJ). They can occupy four positions at most at the same time. Spanish and French predicate verb can take seven grammatical cases (S, SE, DOB, IOB, PP, COMP, CONJ). They can occupy five positions at most at the same time. Currently, 66 VPs peculiar to Spanish are defined in addition to English VPs. Seven VPs peculiar to French are defined, in addition to Spanish VPs.

According to each VP, the following information is revealed.

1. Sets of obligatory grammatical cases
2. Semantic and syntactic constraints for the arguments
3. Mapping information between grammatical cases and deep cases
4. Syntactic features for the argument and the pattern.

In the English system, they are stored at the VP table in the four-slot format. In Spanish and French systems, they are stored at expanded five-slot VP tables. They are used in the conceptual wording process.

3.2 Morphological features for personal pronouns

Table 2 shows personal pronouns and the correspondence between grammatical cases for three languages. This information is stored in each dictionary and is used for word ordering and morphological generation,

3.3 Syntactic and semantic features for a subject

Table 3 shows the features for a subject in three languages, This information is stored in each VP table and is used for subject selection in conceptual wording.

4 Sentence generation model

The sentence generation model consists of four procedures; conceptual wording (sentence-structure planning), syntactic selection, word ordering and morphological generation [2,6]. Among them, conceptual wording is the most important. This section describes an overview of each procedure, and then states the subject and verb pattern selection in the conceptual wording [7].

4.1 Overall sentence generation process

1. Conceptual wording
   This module performs sentence-structure planning by determining the target sentence structure pragmatically and stylistically. So, the language-independent conceptual structure may be transformed into a language-dependent semantic structure, which makes it possible to synthesize a more simple target sentence. In addition to the transformation, this module makes the following decisions.
   (a) Clause category selection : main clause, a subordinate clause, a relative clause, a phrase.
   Which clause/phrase is suitable for the interlingua ?
   (b) Subject and Predicate verb pattern selection.
   What is the subject and predicate in each clause ?
   Which verb pattern is the most suitable for the structure ?

   Conceptual wording rules are mostly shared in the three languages,
2. Syntactic selection

Table 1: Sentence structure and components

		English	French	Spanish
Sentence Structure	Components	Subject (S), Complement (COMP), Indirect Object (IOB), Direct Object (DOB), Prepositional Phrase (PP), Conjunction (CONJ)		
		Adverbial Particle (ADVP)	Reflexive Pronoun (SE)	
	Order	S V IOB DOB PP COMP CONJ	S SE V DOB IOB PP COMP CONJ	S SE V DOB IOB PP COMP CONJ

Table 2: Morphological features for personal pronouns

		English	French	Spanish
Personal Pronoun	Case	Nominative, Possessive, Reflexive		
	Oblique		Indirect Object(Dative),Direct Object(Accusative), Ablative	
	Order		S SE Dative Accusative Verb	
Grammatical Case and Personal Pronoun	SUB	Nominative		
	IOB	Oblique	Dative, Preposition+Ablative	
	DOB		Accusative,Preposition + Ablative	
	PP		Ablative	
	SE		Reflexive	

Table 3: Syntactic and semantic features for a subject

		English	French	Spanish
Subject	Definition	Subject dominates an inflection for the predicate.		
	Position	Subject is located on the left of a verb in the standard position.		Subject can take the right position on some verbs
	Agreement	Subject controls number and gender agreement for a predicate verb.		
		No agreement in Adjective	Subject dominates an inflection of the complement; predicative adjective, past participle and noun.	
	Nominative Pronoun	Subject is not omitted.		Subject can be omitted.
	Impersonal Subject	English "it" and French "il" are used.		Omitted.
		Climate and Time expression.		
		Distance expression		Existence expression.
				Subject is a general person or ambiguous.
	Grammatical case which can be a subject by passivization	Direct Object		
		Indirect Object Prepositional Phrase	Indirect Object	
				Impersonal, Impersonal reflexive and Reflexive are used, instead of passive voice.

The syntactic selection module determines the syntactic information regarding each node in the semantic structure. It also produces the morphological information for surface cases, modals, voice, and so on. As a result, a syntactic structure is created. In this stage, word-order is not settled on.

Syntactic selection rules are similar in French and English, and a little different in Spanish. The different information is described in each VP table.

3. Word ordering

This module determines word-order properties on the syntactic structure. In this way, syntactic generation is divided into two stages; syntactic selection and word ordering. Owing to these two independent determinations, common rules are applicable to the similar syntactic phenomena,

Word ordering rules are mostly similar for English and French, and a little different in Spanish.

4. Morphological generation

Morphological generation uses the word-order properties to arrange the nodes for the grammatical structure in a sequential order, and then generates surface morphemes for each node. Finally, the morphemes are combined into words.

Morphological differences are described as lexical information. Generation rules are almost similar in Spanish and English, and a little different in French.

4.2 Subject selection

Subject selection is to map a deep case onto a grammatical case "Subject" with VP table and the key information regarding the pragmatic function. In this stage, deep cases, which were desirable to be used as the subject, were selected, but were not always unique.

1. Collecting of deep cases as subject candidates

   The following deep cases are collected.

   Those represented in the interlingua.

   Those omitted from the interlingua, but known to the listener.

2. Limitation by dictionary information.

   Using the dictionary information for the predicate verb, deep cases which cannot become a subject, were removed from the collected set.

3. Selection from the collected set, according to the following viewpoints.

   A deep case, which corresponds to a subject case in the source language.

   A deep case, which plays a theme role in the source language.

   A deep case, which exists in interlingua and can be a subject.

   A deep case, which doesn't exist in interlingua, but can still be a subject.

In this subject selection process, pragmatic information, such as Topic, Focus, Theme and Position as the argument for a predicate, plays a very important role.

4.3 Verb pattern selection

Verb pattern selection is implemented to determine the sentence structure. Deep cases, under the predicate CP, should be appropriately mapped onto the grammatical cases. The most suitable verb pattern is selected. Using VP table and the following conditions, VP plausibility could be calculated. Then, the VP with the highest priority was selected as an appropriate one.

1. The semantic features for dependent CPs restrict the possible number of VPs for a predicate CP. VP priority is decreased by this cooccurrence restriction,
2. The priority for VPs, whose subject is inconsistent with the selected subject is increased.
3. The priority for VPs, which can map the greatest number of deep cases onto grammatical cases, is increased.
4. The priority is calculated by the argument position (before or after the predicate) and new/old information for the arguments.

5 Conclusion

The authors comparatively analyzed subjects, sentence structures, grammatical cases, voice and morphological features in English, Spanish and French, and found it possible to develop Spanish and French sentence generation systems by improving the English generation system. This strategy saved much effort for their development, and lead to the multi-lingual sentence generation system with a dictionary of 70,000 English words, 5,000 Spanish words, 3,000 French words, 110,000 Japanese words and 10,000 Korean words.

The authors will continue Spanish and French dictionary development in order to confirm the effectiveness of this strategy.

References

[1] K.Muraki "VENUS: Two-phase Machine Translation System," *Future Generations Computer Systems,* 2, pp.117-119, 1986

[2] K.Muraki "Conceptual Dependency Structure and English Sentence Generation," *Proc. WGNLC of the IEICE,* NLC44-3, July 1984 (in Japanese).

[3] S.Ichiyama "Multi-lingual Machine Translation System," *Office Equipment and Products,* 18-131, pp.46-48, August 1989

[4] C.H.Kim, B.H.Choe, G.C.Kim, K.Choi, S.Ichiyama, "Generation of Korean from Conceptual Representation," *Proc. Information Processing Society of Japan ,* 2B-8, pp. 947-948, October 1986

[5] H.Suematsu "A Simultaneous Processing Approach to the Selection of English Sentence Patterns," *Proc. WGNLC of the IECIE*, NLC89-41, May 1990

[6] H.Suematsu "PIVOT J-E: An Overview of English Generation," *Proc. IECIE*, D-141, October 1988 (in Japanese).

[7] Y.Fukumochi "PIVOT: Subject and Predicate Selection in English Generation," *Proc. Information Processing Society of Japan,* 5C-3, pp.1094-1095, October 1984 (in Japanese).

[8] C.Boite "Multilingual Machine Translation does not have to be saved by Interlingua," *MMT'90, Tokyo,* November 1990

[9] A.S.Hornby "Guide to Patterns and Usage in English," *2nd* ed*, Oxford University* Press.