Field Structure and Generation in Transfer-Driven Machine-Translation

Michael Paul   Eiichiro Sumita   Hitoshi Iida
ATR Interpreting Telecommunications Research Laboratories
e-mail: {paul,sumita,iida}@itl.atr.co.jp

1 Introduction

In (Furuse, 1996), a transfer-driven machine translation approach is proposed within an example-based framework, using knowledge about pre-translated training sentences for the translation task. Examining regularities in the structure and grammatical role of constituents in both languages, *transfer equivalences* are defined for corresponding constituent structures, even enabling the handling of structurally dissimilar languages like Japanese and English/German.

Important for the acceptability of a translation is the selection of an appropriate word order for the target sentence. In this paper we propose a field structure approach to English and German clause syntax employing the notion of *topological fields*, where sentence patterns are described as a linearization of structural units. We show that the descriptive ability of this field structure model incorporated into our example-based approach for machine translation between Japanese–English (JE) and Japanese–German (JG) achieves a high translation quality.

2 Transfer Equivalences

The idea of cross-language regularities can efficiently be utilized in an example-based approach to machine translation. Given a large corpus of example translations, source as well as target sentences are analyzed according to their constituent structure. Grammatical features common to specific source and target constituents enable the cross-language identification of corresponding constituents.

Transfer equivalences[1] define the relationship between these constituents in the translation context. For example, a Japanese phrase marked with the particle が is not always translated as the subject of the target sentence like in 1a, but can also be marked as the direct object as in 1b.[2]

1a. $[X_{\text{日本語}}\ ^{\text{が}}\ Y_{\text{難しい}}]$
  $\rightarrow$ $[[\text{SUB}\ X_{\mathit{Japanese}}]\ Y_{\mathit{be\ difficult}}]$

1b. $[X_{\text{日本語}}\ ^{\text{が}}\ Y_{\text{話せる}}]$
  $\rightarrow$ $[[\text{OBJ}\ X_{\mathit{Japanese}}]\ Y_{\mathit{can\ speak}}]$

Additionally, transfer equivalences define linear precedence constraints for the respective substructures in the context of the example translation. For example, the phrase 他のホテル in 2a is translated as the noun phrase "another hotel", where the adjectival modifier has to be placed in front of the verb. However, in the context of 2b the nominal modifier 京都の is translated as the prepositional phrase "in Kyoto" which has to be attached after the modified noun.

2a. $[X_{\text{他}}\ ^{\text{の}}\ Y_{\text{ホテル}}] \rightarrow [\text{NP}\ X_{\mathit{another}}\ Y_{\mathit{hotel}}]$

2b. $[X_{\text{京都}}\ ^{\text{の}}\ Y_{\text{ホテル}}] \rightarrow [\text{NP}\ Y_{\mathit{hotel}}\ X_{\mathit{in\ Kyoto}}]$

The set of transfer equivalences extracted from the example database represents the empirical knowledge about the structural relationship between source and target language.

3 Utilization of Example Translations

The structure of an input sentence is analyzed according to corresponding syntactic and semantic substructures of example sentences. The source constituents of all transfer equivalences that can be applied to the input are matched and the most appropriate ones are selected.

In order to limit the explosion of structural ambiguity during parsing, the transfer equivalences are attached to several linguistic levels whose hierarchical order restricts the search space for appropriate matches. Thus, the transfer equivalence in 3 which is defined on sentence level is not applied to the analysis of noun phrase structures as in 1a and 1b.

3. $[X_{\text{話せる}}\ ^{\text{が}}\ Y_{\text{読めない}}]$
  $\rightarrow$ $[[\text{SEN}\ X_{\mathit{can\ speak}}]\ [\text{SEN but}\ Y_{\mathit{can't\ read}}]]$

Semantic ambiguities as in 1a and 1b, are resolved by selecting the most appropriate transfer equivalence based on semantic distance calculation, i.e. the distance between the semantic attributes of the respective example and the input words within a thesaurus component (Sumita, 1992).

The combination of analyzed substructures according to the hierarchical order of the selected transfer equivalences determines the source structure of the input sentence.

The translation of the analyzed source structure is carried out in the context of example translations, i.e. the transfer equivalences of the selected examples are applied to the respective source constituents. The corresponding target constituents exhibit the same grammatical role and order constraints as the target constituent of the chosen example translation.

[1] cf. (Kinoshita et al., 1992) for the similar concept of *translation equivalences*

[2] The constituent variable X_{word} defines the respective subconstituent with head *word*. The superior elements separating constituent variables mark a constituent boundary.

This alignment procedure is illustrated for the Japanese utterance in 4.

4. “福岡の他のホテルがありませんか”
⇓
“Is there another hotel in Fukuoka?”

Given the transfer equivalences 2a and 2b, the phrase 福岡の他のホテル can be analyzed as [福岡 $^{\text{の}}$ [他 $^{\text{の}}$ ホテル]] by first matching 2a against 他のホテル and then matching 2b against 福岡の [他 $^{\text{の}}$ ホテル]. This phrase is marked with the particle が. Thus, 1a and 1b could be applied to the sentence structure [X$_{\text{ホテル}}$ $^{\text{が}}$ Y$_{\text{ある}}$]. Because the semantic distance between ある and 話せる is closer than between ある and 難しい, 2b is selected. Finally, an appropriate transfer equivalence [X$_{\text{あり}}$ $^{\text{ません か}}$] matching the sentence predicate is selected resulting in the following source structure:

5. [[[福岡 $^{\text{の}}$ [他 $^{\text{の}}$ ホテル]] $^{\text{が}}$ [あり]] $^{\text{ません か}}$]

In accordance with the selected examples, the transfer knowledge of the respective transfer equivalence is applied to each constituent. The source constituent [他 $^{\text{の}}$ ホテル] is translated as the noun phrase [$_{NP}$ another hotel] where the linear precedence constraint defined in 2a achieves the correct word order within the phrase. Due to the transfer equivalence 2b the prepositional phrase [$_{PP}$ in Fukuoka] modifies the noun “hotel” and is placed after the noun phrase. The source constituent [X$_{\text{ホテル}}$ $^{\text{が}}$] is matched by the transfer equivalence 1b. Thus, the corresponding target constituent is marked as the direct object of the target predicate. Additionally, in the context of the selected transfer equivalence [X$_{\text{あり}}$ $^{\text{ません か}}$], the target sentence is analyzed as a yes/no-question without negation resulting in the target structure 6.

6. [YN-Q [$_{SUB}$ there] [$_{VP}$ be]
[$_{OBJ}$ [$_{NP}$ [$_{NP}$ another hotel] [$_{PP}$ in Fukuoka]]]]

The information of target sentence constituents and their grammatical role is then used to generate an appropriate translation.

4 Evaluation of Word Order

Word order variations occur to some degree in all natural languages. In contrast to more configurational languages like English, languages with partially free word order like German allow numerous variations, where the acceptability of a translation depends on the selection of certain word orders. In the worst case, word order variations can even lead to ungrammatical and incomprehensible sentences.

In the current framework, the handling of word order is limited to the specification of linear precedence constraints in the definition of transfer equivalences. The precedence rules order sister constitutents relative to each other without taking the overall sentence structure into account. Thus, the local scope of transfer equivalences prevents the analysis of order regularities between target constituents at sentence level.

In the following section we define field structure models for the clause syntax of English and German, which exploit syntagmatic regularities of the respective language and show how these models can be utilized in connection with analyzed sentence constituent structure to account for word order variations permitted within the respective clause syntax.

5 Field Structure

In the theory of topologocial fields (Höhle, 1986), clause syntax is described in terms of classes of adjacent word clusters. A *topological field* is an area within the clause that can admit certain constituents. Sentence patterns are defined according to the linearization of topological fields.

As pointed out in (Ahrenberg, 1990), field structure is not exhibited by all languages, but fixed positions of certain sentence constituents might be significant for field structure languages.

5.1 English

The English clause syntax is characterized by a fixed position of the subject. We can distinguish three sentence patterns: questions (*Sq*), statements (*Ss*) and subordinated clauses (*Sc*).

7.

Ss	D_L	–	SUB	VC	VCOMP	POST
Sq	D_L	PRE	SUB	VC	VCOMP	POST
Sc	C		SUB	VC	VCOMP	POST

In all three sentence patterns the *SUB* field is succeeded by a verb complex (*VC*) and the complements of the sentence predicate (*VCOMP*).

The *POST* field is used for the extraposition of sentence elements which otherwise would occur clause-internal, but it also represents the canonical position for placing sentential complements.

Statements as well as questions might be preceded by sentence elements of left dislocation constructions, which are placed in the D_L field.

In contrast to statements, subjects in questions are immediately preceded by an auxiliary verb and in the case of complementary questions by a wh-phrase. These are placed in the *PRE* field.

Elements introducing a verb-final clause, e.g. a complementizer or fronted constituents in relative clauses have to occupy the *C* field.

5.2 German

In German, the clause structure depends on the position of the *finite verb*, i.e. the verbal part of the sentence predicate that is inflected for tense and mood as well as for number and person in agreement with the sentence subject. The placement of the finite verb

Table 1: English Sentence Patterns

	D_L	PRE		SUB	VC				VCOMP				POST
Sm	left	–		sub	aux_1	mod	aux_m	verb	io	do	comp	po	right
Sq	left	wh	aux_f										
Ss	C												

Table 2: German Sentence Patterns

	D_L	PRE	FIN	MID						VC		POST
V1	left	–	vfin	sub/ rpron	mod	obj	neg	unm	po	vnfin		right
V2	left	sub		rpron								
Vf	C			sub/ rpron						vnfin	vfin	

plus the non-finite parts of the sentence predicate forms a skeleton that defines a topological structure for German, which can be divided into the three major sentence patterns listed in 8.

8.

V1	D_L	–	FIN	MID	VC	POST
V2	D_L	PRE	FIN	MID	VC	POST
Vf	C			MID	VC	POST

Verb-first sentences (*V1*) are characterized by the frontal position of the finite verb in the *FIN* field.

In the *verb-second* case (*V2*) the finite verb occupies the same field, but it is preceded by a sentence constituent. Both the *V1* and *V2* sentence pattern might be preceded by sentence elements of left dislocation constructions (D_L). The non-finite parts of the predicate will occupy the verb complex field (*VC*).

The *verb-final* pattern (*Vf*) differs from the previous two in that the finite verb occurs towards the end of the sentence. It is placed in the *VC* field forming a compound verbal complex together with its governed verbs. Similar to English, the fronted constituents of subordinated sentence are placed in the *C* field.

Common to all three sentence patterns is the adjacent field structure *MID*≪*VC*≪*POST*, where the *MID* field is in some sense the default field for the placement of non-verbal and non-dislocated elements like verb complements.

6 Incorporation of Field Structure

The field structure of each sentence pattern has a direct influence on the linearization of sentence constituents assigned to the respective fields. However, the decision about the elements that can occur, their internal structure, and the partial word order within a particular field is independent from the selection of a specific sentence pattern. These factors depend on grammatical and pragmatical constraints.

In order to incorporate field structure into the current framework, we revise the sentence patterns of the respective languages by defining "positions" for each topological field.

A *position* is defined as a subfield without an internal structure in terms of further subfields. These positions are filled up with sentence constituents according to their grammatical and pragmatical usage.

In order to avoid grammatical incorrectness, we define a *canonical* order for the positions of each sentence pattern. A set of linear precedence rules determines the most general order of sentence constituents. The canonical order adapts not only the field structure properties of the respective pattern, but also specifies generally accepted order relations between sentence elements. For example in English and German, the sentence subject is generally placed in front of object complements, whereas a prepositional object tends to occur towards the end of a sentence.

6.1 English

The field structure of English defined in 7 is refined in table 1 by subdividing the *VC* field into positions for possibly multiple auxiliary verbs (aux_1, aux_m), adverbial modifiers (*mod*) and a regular verb (*verb*). The complement field *VCOMP* is split up in positions for the indirect (*io*), direct (*do*) and prepositional (*po*) objects as well as verbal complements (*comp*).

Additionally, the *PRE* field of the *Sq* sentence pattern is divided into two positions for wh-phrases (*wh*) and the finite auxiliary verb (aux_f).

6.2 German

The refined sentence patterns for German are listed in table 2. The *PRE* field of the *V2* sentence pattern is filled by exactly one constituent, which is in general the subject of the sentence (*sub*).

The *MID* field contains positions for the subject (*sub*), the reflexive pronoun (*rpron*), verb complements (*obj*), adverbial modifiers (*mod*), negation (*neg*) and prepositional objects (*po*). The *unm* position is filled with any constituent that is unmarked for their grammatical function.

The non-finite parts of the sentence predicates are assigned to the *vnfin* position. In the *verb-final* case, the *VC* field is subdivided into two positions, where the canonical position of the finite verb is after the *vnfin* position.

6.3 Determination of Word Order

In our example-based framework, the information about constituent structure and the grammatical role of each constituent is encoded in the transfer structure. The word order at phrase level, i.e. the internal structure of each sentence constituent, is determined by the linear precedence constraints defined in the transfer equivalences applied. At sentence-level, the analysis of the clause type triggers the instantiation of a specific sentence pattern. The main constituents of the target sentence are assigned to positions in accordance with their grammatical function, where the linear precedence constraints defined between positions determine the final word order in the target sentence. Thus, the example in 6 will be generated as:

9.	aux_f	*sub*	*do*	*right*
	Is	there	another hotel in Fukuoka	?

In English, the identification of verb complements is encoded mainly topologically. Word order variations are limited to positions that allow the assignment of multiple constituents, e.g. the specification of multiple prepositional objects to the *po* field. In the current framework, these order ambiguities are resolved in the translation context, i.e. with respect to their relative order in the transfer structure.

In German, various order alterations are possible. However the word order is not completely free, but restricted with respect to sentence topology as well as grammatical constraints. Alternations of the canonical order are achieved by means of additional linear precedence rules.

Of all the fields, the *MID* field is subject to most order alterations between sentence elements. The grammatical function of German object complements is marked morphologically using case attributes. We can distinguish nominative (*nom*), genitive (*gen*), dative (*dat*) and accussative (*acc*) complements. We define the canonical order between these complements according to their case attributes as $nom \ll dat \ll acc \ll gen$.

Furthermore, the syntactic category of the respective complements, e.g. pronominal vs. nominal constituents, has a great influence on the word order in the *MID* field, as the order alterations differ according to their acceptability. For example, it is generally accepted that nominal dative complements precede nominal accussative ones, which agrees with our canonical word order. However, the linear order constraint $dat \ll acc$ seems to be less acceptable, if the accussative complement contains a pronominal constituent. Moreover, a pronominal subject influences the word order by forcing a specified reflexive pronoun to succeed the subject position, whereas in the nominal case it might be placed in front of the subject depending on its contextual usage.

Therefore, the incorporation of precedence constraints based on the syntactic category of constituents, like $acc_{pron} \ll dat_{nom}$ or $sub_{pron} \ll rpron$, improves the acceptability of the translation output.

7 Evaluation

For the evaluation of the applicability of the proposed approach, we extracted a set of 9751 transfer equivalences for JE (3435 sentences/156 dialogs) and a set of 6975 transfer equivalences for JG (2121/112) from training dialogs within a travel conversation domain.

In order to evaluate the accuracy of the structural analysis as well as the overall translation quality, we translated unseen data (1225/69) of the same domain.

The system could achieve structural accuracy of 76.7% for JE and 70.6% for JG respectively, where the difference is due to the smaller amount of transfer equivalences in JG.

However, not all of the correctly parsed sentences could be translated appropriately (9.2% for JE and 10.9% for JG). But, an acceptable translation could be achieved for 8.4% (JE) and 9.2% (JG) of the sentences despite an incorrect parse, resulting in a translation quality of 75.9% for JE and 68.9% for JG.

Comparing the difference between accuracy and quality, we see that field structure applied to both target languages achieves a similar result in spite of more complex grammatical constraints in German.

8 Conclusion

In this paper we examined the applicability of field structure models to the handling of word order for configurational languages like English as well as for languages with partial free word order like German in an example-based approach to machine translation.

We showed that word order variations within the translation context can be addressed efficiently for both language pairs by combining the advantages of an example-based alignment method with the descriptive power of a field structure model resulting in a high translation quality.

References

L. Ahrenberg. 1990. Topological Fields and Word Order Constraints. Edinburgh Working Papers in *Cognitive Science 6*, 1-19.

O. Furuse, and H. Iida. 1996. Incremental Translation Utilizing Constituent Boundary Patterns. In *Proc. of Coling '96*, 412–417.

T. Höhle. 1986. Der Begriff "Mittelfeld", Anmerkung über die Theorie der topologischen Felder. In *Akten des 7.Internationalen Germanisten-Kongresses, Göttingen*, Band 3, 329-340.

S. Kinoshita, J. Phillips and J. Tsujii. 1992. Interaction between Structural Changes in Machine Translation. In *Proc. of Coling '92*, 679–685.

E. Sumita and H. Iida. 1992. Example-Based Transfer of Japanese Adnominal Particles into English. *IEICE Transactions on Information and Systems*, E75-D, No.4, 585–594.