Language, Information and Computation (PACLIC12), 18-20 Feb, 1998, 334-339

An Investigation into the use of Argument Structure and Lexical Mapping Theory for Machine Translation

Shun Ha Sylvia Wong[†] *Peter Hancox*[†]
University of Birmingham

Lexical Functional Grammar (LFG) has been quite widely used as the linguistic backbone for recent Machine Translation (MT) systems. The relative order-free functional structure (f-structure) in LFG is believed to provide a suitable medium for performing source-to-target language transfer in a transfer-based MT system. However, the linguistic information captured by traditional f-structure is syntax-based, which makes it relatively language-dependent and thus inadequate to handle the mapping between different languages. Problems are found in the lexical selection and in the transfer from some English passive sentences to Chinese. The recent development of the relatively language-independent argument structure (a-structure) and the lexical mapping theory in the LFG formalism seems to provide a solution to these problems. This paper shows how this can be done and evaluates the effectiveness of the use of a-structures for MT.

1. INTRODUCTION

LFG [2] has been regarded as a suitable linguistic formalism for transfer-based MT systems. Traditional LFG framework is syntax-based which, as illustrated in Figure 1, represents the syntax structure of a sentence in a

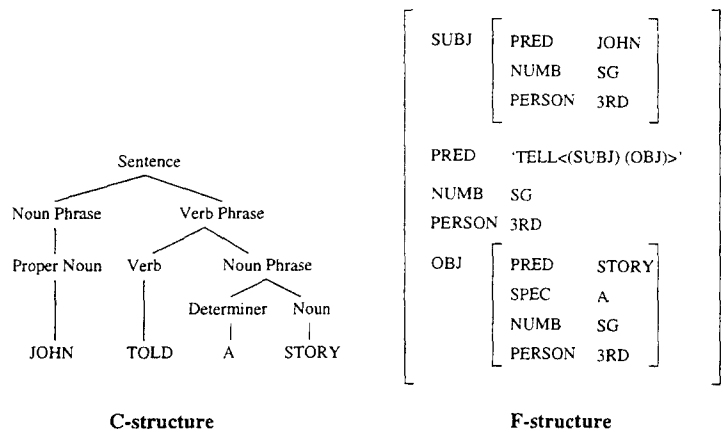

Figure 1: The c- and f-structures for the sentence "John told a story."

hierarchical, tree-like manner (i.e. c-structure) and the higher syntactic and functional information in a relatively order-free functional structure (f-structure). F-structures display linguistic information as relatively order-free attribute-value bundles. This allows linguistic information to be retrieved from or inserted into an f-structure easily for aiding lexical selection during the source-to-target language transfer[1].

Although f-structure provides a suitable medium for transfer, the linguistic information captured in it is syntax-based. Thus, on its own, it is incapable of providing adequate information for word sense disambiguation during the lexical selection. A higher level of linguistic information, which is more language-independent (e.g. semantic information), is required to disambiguate the source language words. However, as traditional f-structures deal with syntactic information only, in the early LFG formalism, there were no guidelines to govern the incorporation of any higher level linguistic information. This makes the use of the LFG formalism for MT less desirable.

[†] School of Computer Science, University of Birmingham, Edgbaston, Birmingham B15 2TT, United Kingdom.
E-mail: S.Wong@cs.bham.ac.uk, P.J.Hancox@cs.bham.ac.uk

[1] The source-to-target language transfer is the process in which source language words are mapped to their corresponding target language forms.

With the aim of improving the ability of the LFG formalism to act as a Universal Grammar for language comparison, recent research on LFG has moved to the extension of the existing structural representation of syntactic and functional information to include some level of semantic information. Recent work on LFG shown that argument structure (a-structure), which represents thematic information of sentences, is capable of capturing more language-independent linguistic information for generalising the similarities across languages [4, 3, 5]. Thematic information represented in a-structures can be incorporated in traditional f-structures according to the lexical mapping theory [5, 6, 12] for enriching the information expressive power of f-structures. This seems to provide a solution for improving the ability of the LFG formalism for MT. The rest of this paper shows how a-structure improves the lexical selection process and how it can solve the problem in transferring some English passive sentences to Chinese.

2. A-STRUCTURE

The participants in an event[2] form the structure of the event. The part taken by each participant in an event is described as *thematic role*. A-structure shows the thematic role played by each participant of the event in each event structure. For instance, the thematic roles which form the event "John told a story." formed the a-structure:

(1) tell<agent theme>

The arguments within the angled brackets describe the thematic roles played by the noun phrases (NPs): 'John' and 'a story' respectively. The thematic roles 'agent' and 'theme' are the least required participants for characterising this event. If the NPs in a sentence cannot be mapped with these thematic roles, it is either describing a different event structure, or the sentence itself is ill-formed. The order of the thematic roles specified within an a-structure corresponds to the thematic hierarchy:

(2) agent < beneficiary < recipient/experiencer < instrument < patient/theme < locative

which reflects the relative prominence of thematic roles characterised by a verb [5, 6]. Although the order of thematic roles within an a-structure does not always reflect the order of the corresponding NPs within a sentence, these orders often agree with each other. Thus, in some cases, the thematic hierarchy helps the mapping of thematic roles within an a-structure to the corresponding NPs within a sentence[3].

3. THE USE OF A-STRUCTURES FOR LEXICAL SELECTION

Most English verbs, when used in different situations, possess different meanings. Though some of these meaning differences are insignificant, when the verbs are translated to Chinese these minute differences can affect the readability of the output translation. Her *et al.* [11] uses the information in semantic forms[4] of verbs to aid lexical selection. However, this kind of information is too syntax-oriented, thus it is insufficient to differentiate the relatively insignificant meaning differences. Carlson [7] pointed out:

> *... verbs assigning different thematic roles should be considered as meaning somewhat different things.*

As thematic roles help to characterise the meaning of verbs, different combinations of thematic roles can, to a certain extent, aid the disambiguation of verbs during the lexical selection process. We used various English verbs and their corresponding Chinese translations in different cases to study the feasibility of using thematic information to differentiate the various meanings possessed by a verb. We found that the use of a-structures, to some extent, is capable of aiding the selection of the most appropriate target translation in MT by differentiating the meaning of the verb used in different cases.

Consider the following sentences:

(3)

English sentence :	John told a story.		English sentence :	I told you!
Chinese translation :	John說了一個故事。		Chinese translation :	我告訴了你！

Though the verb 'tell' is used in both sentences, it is translated as different verbs in Chinese: '說' and '告訴'. The meanings of these Chinese verbs are '*to utter*' and '*to deliver information to someone*' respectively. This meaning difference cannot be distinguished by the semantic form of 'tell' (i.e. 'TELL<(↑ SUBJ) (↑ OBJ)>'), as

[2] An event can be a single action, a state or a process characterised by a verb.

[3] cf. Section 4

[4] A semantic form in traditional LFG framework describes the semantic interpretation of a predicate by the syntactic functions it governs, e.g. the semantic form for the ditransitive verb '*tell*' is 'TELL<(↑ SUBJ) (↑ OBJ2) (↑ OBJ)>' [2].

both of the above usages of 'tell' govern the same syntactic functions: subject and object. However, as suggested by the above meanings (i.e. with or without an explicit recipient of the information in the event), this difference is captured by the different thematic roles assigned for each case:

- 說<agent theme>
- 告訴<agent recipient theme>

Different a-structures are assigned to the verb 'tell' in the above sentences. Thus, the use of a-structures is capable of distinguishing the different senses of 'tell':

- tell<agent theme> = 說<agent theme>
- tell<agent recipient> = 告訴<agent recipient>

As a-structure describes the participants of each event, if the same verb is used to describe different but similar events where their difference lies in the different participant(s) involved, e.g. the verb 'tell' in the above example, the use of a-structure will be more effective in aiding lexical selection than semantic forms.

4. LEXICAL MAPPING THEORY (LMT)

A-structures represent thematic information of sentences which can be used to form a link between lexical semantics and syntactic structures [4]. *Lexical mapping theory* defines how this link can be established by mapping each thematic role within an a-structure to one, and only one, syntactic function of a sentence. This mapping is based on matching some linguistic features possessed by the syntactic functions and thematic roles. These features are [±r] and [±o], where '**r**' stands for *thematically restricted* and '**o**' stands for *objective*. The feature [±r] denotes whether or not the thematic role of a particular syntactic function is fixed, whereas [±o] indicates whether or not a thematic role appears in a sentence as an object. Syntactic functions can be categorised by the features [5, 6, 12]:

$\begin{bmatrix}-r\\-o\end{bmatrix}$ *subject* (SUBJ) $\quad$ $\begin{bmatrix}-r\\+o\end{bmatrix}$ *object* (OBJ)

$\begin{bmatrix}+r\\-o\end{bmatrix}$ *oblique function* (OBL_θ) $\quad$ $\begin{bmatrix}+r\\+o\end{bmatrix}$ $\textit{object}_\theta$ (OBJ_θ)

Some thematic roles possess some of the above features intrinsically. The thematic roles *agent*, *theme* and *locative* possess the intrinsic feature: [-o], [-r] and [-o] respectively. The assignment of additional features to each thematic role within an a-structure is based on [5, Pages 78–79]:

- the morphological operation '*passive*',
- the default feature classification, and
- the well-formedness conditions

With these feature assignment criteria and the information about the intrinsic possession of the [±r] and [±o] features, each thematic role within an a-structure can be associated with the corresponding syntactic function within a sentence by matching the features of the thematic role with that of the most appropriate syntactic function. During feature matching, the system always aims at assigning the thematic role to the syntactic function which possesses exactly the same features. However, if this complete match cannot be carried out, the system will then use the thematic hierarchy and the feature assigned to each thematic role to perform a partial match with the features possessed by the syntactic functions so as to select the most appropriate syntactic function for lexical mapping. At the end of the matching process, according to the well-formedness conditions, each thematic role in the a-structure should be mapped to one, and only one, syntactic function in the sentence; and vice versa. No thematic role within an a-structure or no syntactic function in a sentence should be left unmapped. For instance, the lexical mapping for the English sentence "Mary was given a book by John." is:

(4) Sentence : Mary was given a book by John.

A-structure :	*give* <	agent	recipient	theme >	by <	agent >
Intrinsic :		[-o]		[-r]		[-o]
Passive :	be	∅				
Default :			[+r]			[+r]
Syntactic Functions :			SUBJ	VCOMP OBJ		VCOMP OBL_θ
NPs :			Mary	a book		John

5. THE TRANSFER FROM ENGLISH PASSIVE SENTENCES TO CHINESE

As mentioned earlier, the attribute-value bundle representation of f-structure provides a suitable medium for source-to-target language transfer. Within an f-structure, the linguistic information of a sentence is represented as attribute-value pairs[5]. The attribute-value pairs belonging to the same syntactic function are grouped together[6]. This allows the transfer from a source language sentence to the required target language to be carried out at phrase (or even word) level. The output of the transfer is then assembled to form the required target sentence in the target language sentence generation process. Due to the difference between the source and target grammars, some words in the source sentence are ignored in the transfer process or extra target language words are required to add to the target sentence. Carrying out the transfer at phrase level allows these to be done easily. By breaking down the source sentence into small chunks for transfer makes the whole MT process simpler and easier to manage. However, in order to perform this kind of transfer successfully, the f-structures of the source language sentence and its target language equivalent must have similar hierarchical structure, otherwise it will be difficult to map the source language words and phrases to their corresponding target language form. As traditional f-structures deal with the syntax-oriented information of sentences, they are quite language-dependent. The f-structure of a sentence in one language does not necessarily be identical to that of its target equivalence. We found that the f-structures of English passive sentences and their Chinese counterparts are dissimilar in some ways. As a result, f-structure cannot be used as the sole medium for the transfer. Some *transformation rules* are required to form the target f-structure from the source f-structure for the later target sentence generation process. However, these kind of transformation rules are not defined in the traditional LFG framework.

Consider the grammatical correctness of the following sentences (cf. [12, P.359]):

(5) English sentence (*grammatical*) : Mary was given a book by John.
Chinese translation (*ungrammatical*) : Mary被John送了一本書。

(6) Chinese sentence (*grammatical*) : 一本書被John送了Mary.
English translation (*ungrammatical*) : A book was given Mary by John.

The sentence structure between the English passive sentence with 'give' and its Chinese counterpart '送' are different. The correct translation for the English sentence in (5) is the Chinese sentence in (6). According to Huang [12], the difference between thematic hierarchies for Chinese and English[7] accounts for this structural difference. Even though the Chinese passive marker '被' in (6) functions similarly as the English passive marker 'be', they are different in some ways [10]. As a result, the f-structures of the English sentence in (5) and its Chinese counterpart are different. However, this is not accounted for in the traditional LFG framework. Huang suggested that this difference is shown in the a-structures for 'give' and '送' [12]:

(7) English sentence : Mary was given a book by John.
A-structure : give<agent *recipient theme*>
A-structure : 送<agent *theme recipient*>
Chinese translation : 一本書被John送了Mary.

The order of thematic roles within the a-structures in (7) reflects the order of the correpsonding NPs appears in the passive sentences, i.e. the recipient in a Chinese passive sentence is preceded by the theme. These a-structures can be used to bridge the gap between the Chinese and the English passive sentences. During the transfer, the selection of the most appropriate Chinese verb can be done by matching the thematic roles it possesses with that of 'give'. The order of thematic roles are neglected in this matching process. Due to the different syntactic structures English and Chinese passive sentences possesses, an NP in the English sentence cannot always be mapped to the same syntactic function in the Chinese sentence (or vice verser). To solve this problem, before each syntactic function in the source sentence is transferred to its target equivalent, it is associated with the appropriate syntactic function in the target sentence by the assigned thematic role. As stated in Section 4, the assignment of thematic roles to the appropriate syntactic functions is governed by the lexical mapping theory, the syntactic functions in the source sentence can be associated with that of the target sentence as follows[8]:

[5] An attribute can be a syntactic function or a grammatical feature (e.g. NUM, TENSE). The value for each attribute can be a simple symbol; a semantic form or a subsidiary f-structure [2, pages 176–177].
[6] cf. Figure 1 in Section 1
[7] The thematic hierarchy for Chinese is: agent < beneficiary/maleficiary < instrument < patient/theme < experiencer/goal < locative/domain [12, P. 353]. The difference lies between the order of the thematic roles 'patient/theme' and 'experiencer/goal (i.e. recipient)' (*cf. Section 2*).
[8] cf. Section 4

(8) Source sentence : Mary was given a book by John.

Source NPs :			*a book*	*Mary*		*John*	
Source Syntactic Functions :			VCOMP OBJ	SUBJ		VCOMP OBL_θ	
Target A-structure :	送 <	agent	theme	recipient	> 被 <	agent	>
Intrinsic :		[-o]	[-r]			[-o]	
Passive :	被	∅					
Default :				[+r]		[+r]	
Target Syntactic Functions :			SUBJ	XCOMP OBJ_θ		OBL_θ	
Target NPs :			一本書	Mary		John	

Target sentence : 一本書被John送了Mary.

After this mapping, the skeleton for the target f-structure is formed. Each syntactic function in the source sentence can then be transferred easily according to the linguistic information captured in the source f-structure.

6. DISCUSSION

A-structure has two facets. In semantic terms, as thematic roles describe the different means of participating an event, they show some semantic information about the characteristic of each participant of the event. For instance, the *agent* of an event is an animate object as it is the one responsible for initiating the event [9]. In syntactic terms, each a-structure is linked with the syntactic structure of a sentence by assigning each thematic role to the corresponding syntactic function within the sentence. Due to this dual function, a-structure is capable of acting as a link between lexical semantics and syntactic structures [4]. As exemplified in Sections 1 & 5, the linguistic information captured in a traditional f-structure is language-dependent and thus it is insufficient for aiding moderately sophisticated lexical selection and for transferring some kinds of sentences, e.g. passive, from one language to another. As thematic information only shows the different kinds of participants involved in an event, but not the context of the sentence, although a-structure is capable of aiding the lexical selection, it does not provide sufficient information for carrying out highly sophisticated transfer. For instance, the English verb '**break**' which denotes the change-of-state of an object has numerous translations in Chinese depending on the semantic of the participants [14]. Thematic information is inadequate to transfer these kind of words successfully as the same a-structure can be used to describe the different translation in Chinese.

Palmer and Wu suggested the use of selectional restrictions and conceptual primitives for handling the disambiguation of words with one-to-many translations in the target language [14]. An interlingual conceptual lattice is built by merging the hierarchies of conceptual primitives for verb senses in English and Chinese. The lexical selection was performed by calculating the meaning similarity between words within the conceptual lattice and the best translation is selected based on the calculated meaning similarity. This method is particularly useful when the required MT system is not confined to processing a sublanguage only, but a broader coverage of a natural language. However, in order to ensure its effectiveness, a complicated conceptual lattice is required to be built. Unless an automatic or semi-automatic method is used to develop the required conceptual lattice, the large amount of time and human effort required to build the required system will make this method too costly and difficult to be implemented for real-life MT tasks. Though thematic information is inadequate to support this kind of high-level semantic disambiguation, it allows the disambiguation of a wide range of words, whose translations are dictated by their governing thematic roles, to be performed in a relatively less costly and simple way. In addition, it bridges the gap between lexical semantic and syntactic structures, so that both semantic and syntactic information can be captured and used in the whole MT process. Although Palmer and Wu's method support a highly sophisticated lexical selection, as syntactic information is required for the target sentence generation process, additional syntactic analysis is required. This makes the MT process more difficult to maintain. The use of a good linguistic formalism (e.g. LFG) is proven to provide a complete, linguistically sound and easy-to-understand[9] method for MT. The introduction of some semantic information to f-structures can provide more detailed information for improving the transfer. The improved LFG framework provides means to capture both syntactic and thematic information (i.e. c-, f- and a-structures); no additional means is required to aid the translation process. The resulting MT system is relatively easy to implement and to maintain. As a-structure can act as a link between lexical semantics and syntactic structures, additional semantic information can be incorporated to f-structures fairly easily

[9] Linguistic-based MT method is readily understandable by both theoretical and computational linguists.

in the form of additional attribute-value bundles so as to further improve the ability to select the most appropriate target translation. As thematic roles helps to disambiguate verb sense, the amount of different semantic markers required for more sophisticated disambiguation is reduced.

In this approach, a-structure plays a crucial role in the transfer. In order to implement this approach successfully, it is very important to obtain the a-structure(s) for each verb in the lexicon. Although there is no generally accepted guidelines to govern the establishment of a-structures, there is a wide range of literature written about the formation of argument structures and the characteristic of each thematic role, e.g. [13, 9, 8]. With the aid of a good dictionary which shows all the syntactic functions governed by a verb, the use of any set of guidelines, or a combination of guidelines, and the thematic hierarchy can effectively aid the establishment of a-structures for most verbs.

7. CONCLUSION

LFG has been regarded as a suitable linguistic formalism for natural language processing (NLP). However, the linguistic information that traditional LFG framework deals with is insufficient to support a moderately sophisticated transfer in MT. It is shown that with the introduction of thematic information captured in a-structures by the lexical mapping theory in the recent LFG framework, the transfer process can be improve. Although, to certain extent, the use of thematic information is still insufficient to solve the problem of ambiguity in MT, the use of c-, f- and a-structures and the lexical mapping theory provides a relatively easy-to-implement and efficient method for handling the transfer in MT. As the application of a-structure in NLP is a relatively new research area, it is believed that more research on how a-structures can be established can improve the application of a-structure on MT.

References

[1] Alex Alsina. Resultatives: A Joint Operation of Semantic and Syntactic Structures. In *Proceedings of the 1996 LFG Conference and Workshops*, Rank Xerox, Grenoble, Aug 1996.

[2] Joan Bresnan, editor. *The Mental Representation of Grammatical Relations*. MIT Press, Massachusetts and England, 1982.

[3] Joan Bresnan. Locative Inversion and Universal Grammar. *Language*, 70(1):72–131, 1994.

[4] Joan Bresnan. Lexicality and Argument Structure. In *Syntax and Semantics: Proceedings of a conference*, Paris, Oct 1995.

[5] Joan Bresnan and Jonni M. Kanerva. Locative Inversion in Chichewa: A Case Study of Factorization in Grammar. *Syntax and Semantics*, 26:53–101, 1992.

[6] Joan Bresnan and Annie Zaenen. Deep Unaccusativity in LFG. In *Grammatical Relations: A Cross-Theoretical Perspective*, pages 45–57. The Center for the Study of Langauge and Information (CSLI), 1990.

[7] Greg N. Carlson. Thematic roles and their role in semantic interpretation. *Linguistics*, 22:259–279, 1984.

[8] David Dowty. Thematic Proto-roles and Argument Selection. *Language*, 67:547–619, 1991.

[9] Talmy Givón. *Syntax: a functional-typological introduction*, volume 1. John Benjamins, Amsterdam/Philadelphia, 1984.

[10] One-Soon Her. An LFG account for Chinese bei sentences. *Journal of the Chinese Language Teachers Association*, 23(3):67–89, 1989.

[11] One-Soon Her, Dan Higinbotham, and Joseph Pentheroudakis. Lexical and idiomatic transfer in machine translation: An LFG approach. In *Research in Humanities Computing*, volume 3, pages 200–216. Oxford University Press, Oxford, 1994.

[12] Chu-Ren Huang. Mandarin Chinese and the Lexical Mapping Theroy — a study of the interaction of morphology and argument changing. *The Bulletin of the Institute of History and Philology*, 62:337–388, 1993.

[13] Ray Jackendoff. *Thematic Relations*, pages 29–46. MIT Press, USA, 1972.

[14] Martha Palmer and Zhibiao Wu. Verb Semantics for English-Chinese Translation. *Machine Translation*, 10:59–92, 1995.