Linking a Grammar to an Ontology

Wai Lok TAM[1], Yusuke MATSUBARA[1], Koiti HASIDA[1], Motoyuki TAKAAI[2], Eiji ARAMAKI[3], and Hiroshi UOZAKI[4]

[1]National Institute of Advanced Industrial Science and Technology
[2]Communication Technology Laboratory, Research and Technology Group, Fuji Xerox Co., Ltd.
[3]Centre for Knowledge Structuring. University of Tokyo
[4]Department of Pathology, School of Medicine, Teikyo University

1 Introduction

This paper is a follow-up to (Hasida et al.2012). It illustrates how a Japanese grammar that comes with classes in an ontology as the semantic representations of the leaves works. A trace of parsing a pathological report and the semantic composition that goes hand in hand with it will be done. We start with giving some background information of the ontology and the linguistic data in section 2. Next we would provide an overview of the components of the grammar in section 3. This is followed by our analysis of two example sentences. Finally, we conclude this paper by telling what we do with our grammar.

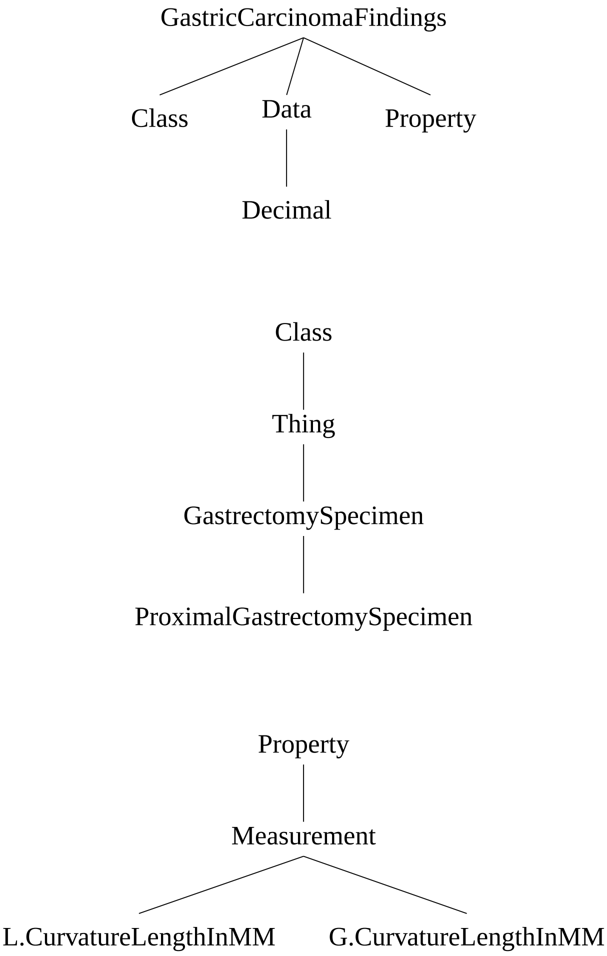
2 Background

Our data are pathological reports of patients diagnosed with stomach cancer. For this paper, we focus our attention on the portion consisting of sentences. Below are two sentences taken from one such portion of a pathological report:

(1) [$_{\text{N}}$ 噴門側胃切除検体]
funmonsokuisetsujyokentai
proximal_gastrectomy_specimen

(2) [$_{\text{N}}$ 小弯長] [$_{\text{NOADJ-GA}}$ 12cm] [, ,]
shouwanchou jyuni_senchi toutenn
[$_{\text{N}}$ 大弯長] [$_{\text{NOADJ-GA}}$ 19.5cm] [。。]
daiwanchou jyukyutengo_senchi kuten
length_of_the_lesser_curvature 12cm ,
length_of_the_greater_curvature 19.5cm .

Next comes a baby version of our ontology trimmed down to concepts relevant to the above sentences (L. is abbreviation for Lesser and G. is abbreviation for Greater).



3 The Grammar

We are still missing the links between concepts in our ontology and words in our source data. These links are given in the lexicon of our grammar:

The links between concepts in our ontology and sentences in our source data are computed by the following semantic composition rules:

Copyright(C) 2013 The Association for Natural Language Processing.
All Rights Reserved.

$$\begin{bmatrix} \text{ORTH} & \textit{funmonsokuisetsujyokentai} \\ \text{POS} & \textit{N} \\ \text{SELF} & \boxed{1}\ \textit{ProximalGastrectomySpecimen} \\ \text{GOV} & \boxed{1} \\ \text{CENT} & \boxed{1} \end{bmatrix}$$

$$\begin{bmatrix} \text{ORTH} & \textit{shouwanchou} \\ \text{POS} & \textit{N} \\ \text{CENT} & \textit{GastrectomySpecimen} \\ \text{SELF} & \boxed{2}\ \textit{decimal} \\ \text{GOV} & \boxed{2} \end{bmatrix} \quad \begin{bmatrix} \text{ORTH} & \textit{daiwanchou} \\ \text{POS} & \textit{N} \\ \text{CENT} & \textit{GastrectomySpecimen} \\ \text{SELF} & \boxed{3}\ \textit{decimal} \\ \text{GOV} & \boxed{3} \end{bmatrix}$$

$$\begin{bmatrix} \text{ORTH} & \textit{12.0cm} \\ \text{POS} & \textit{NOADJ-GA} \\ \text{SELF} & \boxed{4}\ \textit{120} \\ \text{GOV} & \boxed{4} \end{bmatrix} \quad \begin{bmatrix} \text{ORTH} & \textit{19.5cm} \\ \text{POS} & \textit{NOADJ-GA} \\ \text{SELF} & \boxed{5}\ \textit{195} \\ \text{GOV} & \boxed{5} \end{bmatrix}$$

Figure 1: Lexicon

$$\begin{bmatrix} \text{SELF} & \boxed{9} \\ \text{GOV} & \boxed{10} \\ \text{CENT} & \boxed{11} \end{bmatrix}^{M} \rightarrow \begin{bmatrix} \text{SELF} & \boxed{27} \\ \text{GOV} & \boxed{9} \\ \text{CENT} & \boxed{11} \end{bmatrix}^{NH} + \begin{bmatrix} \text{SELF} & \boxed{9} \\ \text{GOV} & \boxed{10} \\ \text{CENT} & \boxed{11} \end{bmatrix}^{HD}$$

where the mother(M), the head daughter (HD) and the nonhead daughter (NH) are determined by the combination of POS labels in the syntactic rules given below:

$$\begin{array}{rcl} NOADJ & \rightarrow & N + \underline{NOADJ - GA} \\ NOADJ+, & \rightarrow & \underline{NOADJ}+, \\ NOADJ., & \rightarrow & \underline{NOADJ+,} + \underline{NOADJ} \\ SP + \circ & \rightarrow & \underline{NOADJ.,} + \circ \end{array}$$

The underlined daughter is the head daughter of the mother on the left hand side. The rule with two head daughters goes hand in hand with the conjunction rule, which requires the GOV and the SELF features to become partial functions:

$$\begin{bmatrix} \text{SELF} & \boxed{12} \\ \text{GOV} & \boxed{13} \\ \text{SELF} & \boxed{14} \\ \text{GOV} & \boxed{15} \\ \text{CENT} & \boxed{24} \end{bmatrix}^{M} \rightarrow \begin{bmatrix} \text{SELF} & \boxed{12} \\ \text{GOV} & \boxed{13} \\ \text{CENT} & \boxed{24} \end{bmatrix}^{HD1} + \begin{bmatrix} \text{SELF} & \boxed{14} \\ \text{GOV} & \boxed{15} \\ \text{CENT} & \boxed{24} \end{bmatrix}^{HD2}$$

Now let us parse the example sentence (2) with the syntactic rules. The parse tree is given in figure 2. The semantic representations of the nodes labelled boxed 18, 19 and 20 are given in figure 3 to show the effect of applying the semantic rule them.

To make sense of the semantic composition going on here, some explanation for the feature names and the values is probably needed. *decimal* is supposed to be a data type and ”120” belongs to this type. So unifying ”120” with *decimal* yields ”120”. The result of unifying a class, whose name begins with an upper case letter, with another class is determined by the ontology. Both classes and data are possible values of semantic features. The semantic features in figure 3 are: SELF, GOV, CENT, and LesserCurvatureInMM. SELF and GOV are fundamental to all constituents except for the period and comma. The former feature is equivalent to the meaning of the constituent itself whereas the later feature, if not structure-shared with the former, is the meaning of the head on which the constituent depends. The last feature, LesserCurvatureInMM, corresponds to a property of the class GastrectomySpecimen defined in the ontology. Such feature may show up as subfeatures of values of SELF, GOV and CENT and relate them to each other.

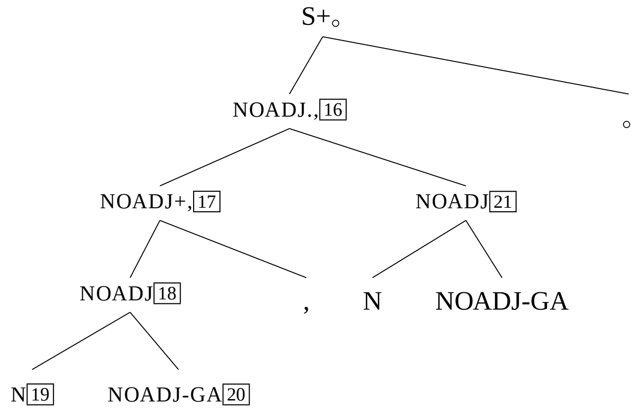


Figure 2: parse tree of example sentence (2)

$$\boxed{18} \begin{bmatrix} \text{ORTH} & \textit{shouwanchou} \\ \text{POS} & \textit{NOADJ} \\ \text{CENT} & \boxed{23}\ \textit{GastrectomySpecimen} \\ \text{SELF} & \boxed{22}\ \textit{decimal} \\ \text{GOV} & \boxed{22} \end{bmatrix},$$

$$\boxed{19} \begin{bmatrix} \text{ORTH} & \textit{shouwanchou} \\ \text{POS} & \textit{N} \\ \text{CENT} & \boxed{23}\ \textit{GastrectomySpecimen} \\ \text{SELF} & \boxed{22}\ \textit{decimal} \\ \text{GOV} & \boxed{22} \end{bmatrix},$$

$$\boxed{20} \begin{bmatrix} \text{ORTH} & \textit{12.0cm} \\ \text{POS} & \textit{NOADJ-GA} \\ \text{SELF} & \boxed{22}\ \textit{120} \\ \text{GOV} & \boxed{22} \end{bmatrix}$$

Figure 3: Node Boxed 18, Boxed 19 and Boxed 20

The LesserCurvatureInMM feature in 4 relates the decimal number ”120” assigned to the SELF feature to an instance of the *GastrectomySpecimen* class assigned to the CENT feature. When the CENT feature is assigned any value other than *Thing*, it indicates a zero anaphora. In a zero anaphoric relationship, the antecedent is not referred to by a constituent but a gap. To resolve such anaphora, there is first a need to put this gap in the representation of a constituent depending on it or a constituent on which it depends if we do not create a node for it in a parse tree. Meeting this need is one of the purpose of the CENT feature. There is also a second need to pass up the CENT value to the root node for the antecedent to access. The first step of this percolation is illustrated in figure 3. The second step is trivial because the sister of node boxed 18 does not get any semantic features (Notice that unifying a

Copyright(C) 2013 The Association for Natural Language Processing.
All Rights Reserved.

$$\boxed{23}\left[\textsc{LesserCurvatureInMM}\ \boxed{22}\ \textit{decimal}\right]$$

Figure 4: Boxed 23 at Node Boxed 18 and Boxed 19

feature value with a non-existing feature value results in the feature value being passed up in our grammar). This means the values of semantic features of node boxed 18 are passed up without any changes to node boxed 17. To further pass up the values of features of node boxed 17, we would have to draw on the conjunction rule. The conjunction rule is applied to node boxed 17 and node boxed 21. The representation of node boxed 21 and the result of rule application, that is, node boxed 16, is given in figure 5.

$$\boxed{21}\begin{bmatrix}\textsc{orth} & \textit{shouwanchou}\\ \textsc{pos} & \textit{NOADJ}\\ \textsc{cent} & \boxed{23}\\ \textsc{self} & \boxed{25}\ \textit{195}\\ \textsc{gov} & \boxed{25}\end{bmatrix},\quad \boxed{16}\begin{bmatrix}\textsc{self} & \boxed{22}\\ \textsc{gov} & \boxed{22}\\ \textsc{self} & \boxed{25}\\ \textsc{gov} & \boxed{25}\\ \textsc{cent} & \boxed{23}\end{bmatrix}$$

Figure 5: Node Boxed 16 and Boxed 21

$$\boxed{23}\begin{bmatrix}\textsc{LesserCurvatureInMM} & \boxed{22}\ \textit{decimal}\\ \textsc{GreatererCurvatureInMM} & \boxed{25}\ \textit{decimal}\end{bmatrix}$$

Figure 6: Boxed 23 at Node Boxed 21

We are now one step from the root node in figure 2. There is not much to say about this step because the sister of node boxed 16 does not get any semantic feature. So the semantic features of the root get exactly the same value as node boxed 16. We are not done at the root node of example sentence (2). We still get the anaphoric gap labelled boxed 23 to be filled. The filling is done by pairing the root node of example sentence (2) with the root node of example sentence (1) containing the antecedent as illustrated by the anaphora resolution rule given below.

$$\begin{bmatrix}\textsc{self} & \boxed{22}\\ \textsc{gov} & \boxed{22}\\ \textsc{self} & \boxed{25}\\ \textsc{gov} & \boxed{25}\\ \textsc{self} & \boxed{23}\\ \textsc{gov} & \boxed{23}\\ \textsc{cent} & \boxed{23}\end{bmatrix} \rightarrow \begin{bmatrix}\textsc{self} & \boxed{23}\\ \textsc{gov} & \boxed{23}\\ \textsc{cent} & \boxed{23}\end{bmatrix} + \begin{bmatrix}\textsc{self} & \boxed{22}\\ \textsc{gov} & \boxed{22}\\ \textsc{self} & \boxed{25}\\ \textsc{gov} & \boxed{25}\\ \textsc{cent} & \boxed{23}\end{bmatrix}$$

4 Conclusion

The grammar described above is the core of an input suggestion system built for writing pathological reports. This makes our solution unique in that we do it correctly in dealing with a world knowledge problem by a world knowledge approach (organizing such knowledge in an ontology and link it to a grammar), when compared to statistical systems that typically just lump everything together.

Acknowledgements

This work is supported by the Adaptable and Seamless Technology Transfer Program through target-driven R&D (Japan Science and Technology Agency).

References

Hasida, Koiti, Wailok Tam, Taiichi Hashimoto, Motoyuki Takaai, and Eiji Aramaki. 2012. Ontoroji taiou bunpou riron to sono kousokushori no tame no conpaireson. In *Proceedings of Gengo Shori Gakkai*, Hiroshima, Japan.

Copyright(C) 2013 The Association for Natural Language Processing.
All Rights Reserved.