Evaluating HILDA in the CODA Project: A Case Study in Question Generation Using Automatic Discourse Analysis

Pascal Kuyten (1) **Hugo Hernault** (1) **Helmut Prendinger,** (2) **Mitsuru Ishizuka** (1)

(1) Graduate School of Information Science & Technology
The University of Tokyo
Tokyo, Japan
{pascal, hugo}@mi.ci.i.u-tokyo.ac.jp
ishizuka@i.u-tokyo.ac.jp

(2) National Institute of Informatics
Tokyo, Japan
helmut@nii.ac.jp

Abstract

Recent studies on question generation identify the need for automatic discourse analysers. We evaluated the feasibility of integrating an available discourse analyser called HILDA for a specific question generation system called CODA; introduce an approach by extracting a discourse corpus from the CODA parallel corpus; and identified future work towards automatic discourse analysis in the domain of question generation.

Question Generation

Question generation is an important and challenging component of systems where knowledge extraction and representation in natural language is desired (Rus et al. 2010). Rus and Graesser defined Question Generation as the task of automatically generating of questions from some form of input. The input could vary from information in a database to a deep semantic representation to raw text (Rus and Graesser 2009).

Studies on question generation can be classified by considering their scoping. Some generate questions at sentence level and others generate questions at paragraph level (Rus et al. 2010). When considering question generation at paragraph level the discourse relations become important (Heilman 2011).

Two studies which are generating questions at paragraph level are the CODA project (Piwek and Stoyanchev 2010a) and the work by Mannem et al. (Mannem, Prasad and Joshi 2010).

The CODA project is a two years program that started in 2009 and is dedicated to generation of dialogue. Question generation is identified as an important component in the generation of dialogue. Dialogue is generated from monologue text using the CODA system. A part of CODA's dialogue generation consists of discourse annotating the monologue text, thereafter applying rules to generate dialogue acts. These rules were extracted from the CODA parallel corpus, of which 70% are based on a discourse relation in the monologue (Piwek and Stoyanchev 2010a).

Mannem et al. introduce a question generation system developed for the question generation challenge (Rus et al. 2010). As current automatic discourse analyzers are in an early stage, a different approach is introduced by applying semantic role labeling. Results indicate that further expansion of the parsing and transformation rules may increase overall performance of question generation using semantic role labeling (Mannem, Prasad and Joshi 2010).

Automatic Discourse Analysis

Several recent studies focus on developing automatic discourse analysers (Soricut and Marcu 2003), (Reitter 2003), (LeThanh, Abeysinghe and Huyck 2004) and (Hernault et al. 2010). Two of these studies provide publicly available implementations: SPADE, based on the work of (Soricut and Marcu 2003), provides sentence level discourse analysis; and HILDA (Hernault et al. 2010), based on Support Vector Machine classification (Hernault, Bollegala and Ishizuka 2011), provides text level discourse analysis.

Challenges in automatic discourse analysis can be considered in three areas. First, discourse segmentation: texts must be accurately segmented into elementary discourse units (EDUs). Second, discourse representation: EDUs' relations need to be represented in some discourse structure. Third, discourse corpora: a class of automatic discourse analysers learn from annotated discourse corpora, of which there are few (Hernault, Bollegala and Ishizuka 2011).

Case Study

Integrating an automatic discourse analyser into the CODA system for the segmentation and annotation of text is noted as future work (Stoyanchev and Piwek 2010b). The CODA system requires paragraph level discourse analysis. HILDA performs text level discourse analysis and SPADE performs sentence level discourse analysis, therefor we chose to evaluate HILDA as a candidate for the automatic discourse analyser for the CODA system. We use the publically available version of HILDA, which has been trained with the RST Discourse Treebank corpus (Hernault et al. 2010).

The CODA parallel corpus has been created for rules' extraction, mapping monologue text into dialogue text, and is based on dialogue text. Dialogue text is manually rewritten into monologue text, thereafter annotators described their discourse.

The CODA's discourse structure may differ from the RST Discourse Treebank. For example CODA's annotators were asked to prioritize discourse relations (Stoyanchev and Piwek 2010b), potentially creating a mismatch between the discourse analysis generated and the discourse analysis expected. Therefor we propose to use the CODA parallel corpus to evaluate the performance of a candidate for the automatic discourse analyser in the CODA system.

Setup

The CODA parallel corpus consists of a collection of paragraphs. Each paragraph includes: a monologue text; the source dialogue; and a manual annotated discourse analysis represented by an RST tree.

First, the monologue texts and RST trees from the CODA parallel corpus are extracted and normalised; Second, each of the monologue text discourse are analysed by HILDA; Third, the obtained RST trees are normalised; Fourth, the CODA's RST trees are compared with the HILDA's RST trees; and Fifth, the comparison is analysed.

Extraction and Normalisation

The monologue text is already separated into segments; these segments are concatenated with the "<s>"- delimiter and stored in a file. The delimiter is a requirement of HILDA and is used for guidance of the segmentation process.

The CODA parallel corpus embodies a different set of discourse relations than HILDA does, because HILDA's discourse relations are more general than CODA's discourse relations, CODA's discourse relations are combined following the two left-most columns of Table 1.

Some of the CODA's RST trees are structured in a non-linear reading order; such trees are normalized into a linear reading order. E.g. if a monologue text consists of the segments $\{S_0, S_1, S_2\}$ and the RST tree is structured as $[R_0: S_2N, [R_1: S_0N, S_1]]$, then the RST tree is normalized to $[R_0: [R_1: S_0N, S_1], S_2N]$, where S describes the segment, R describes the discourse relation and N describes the nucleus.

HILDA normalises the input text, for example by removing whitespaces and changing UK spelling to US spelling; in order to allow comparison all EDUs are normalized.

Comparison and Analysis

The CODA's RST trees are compared with HILDA's RST trees by comparing all sub-trees recursively. A comparable set of two sub-trees are located at the exact same location in their corresponding RST tree. For example: CODA's RST tree's root node is compared with the HILDA's RST tree's root-node, thereafter CODA's RST tree's left child is compared with the HILDA's RST tree's left child etc.

The comparison is analysed in terms of structure: depth and balance; and content: discourse relation (R); EDUs in the satellite and nucleus (T); and position of the nucleus and satellite (O).

Depth of a sub-tree is defined as the number of nodes between sub-tree's root and the furthest child, where both the root-node and child-node are counted. E.g. the sub-tree $[R_0: S_2N, [R_1: S_0N, S_1]]$ has a depth of 3 and $[R_1: S_0N, S_1]$ has a depth of 2.

Balance of a sub-tree is defined as the depth of the left child divided by the depth of the right child. E.g. the sub-tree $[R_0: S_2N, [R_1: S_0N, S_1]]$ has a balance of 0.5.

Results

The CODA parallel corpus contained 192 monologue texts of which 178 could be used for comparison, 6 texts contained dummy segments and 7 texts were annotated with non-comparable RST trees. HILDA generated successfully 142 RST trees and failed for 36 monologue texts. Failure occurred primarily due to segmentation issues.

CODA's RST trees associated with the 142 analysable monologue texts contained 291 sub-trees; HILDA's RST trees contained 1150 sub-trees, of which 248 could be used for comparison. The other sub-trees were present at different locations in the RST tree. From these 248 comparable sub-trees, only 58 sub-trees had matching EDUs in their left and right children.

A comparison of sub-trees is listed in Table 1. The C-column describes the comparable sub-trees; and the RTO-column describes the comparable sub-trees with matching discourse relations; EDUs found in their left and right children; and position of their nucleus and satellite.

Discourse relation		Number			Average depth				Average balance			
CODA	*HILDA*	*CODA*	*HILDA*		*CODA*	*HILDA*			*CODA*	*HILDA*		
			C	All		RTO	**C**	All		RTO	**C**	All
Attribution	Attribution	12	**10**	128	2.2	2.0	4.2	3.5	.9	1.0	**.8**	1.4
Background	Background	3	**9**	29	3.0		5.2	3.5	.8		**1.3**	1.2
Cause	Cause	4	**2**	5	3.3		3.0	2.8	1.1		**2.0**	1.6
Comparison	Comparison	5	**0**	8	2.6			2.5	.9			1.0
Condition	Condition	24	**5**	18	2.3	2.5	2.8	3.1	1.0	1.5	**1.5**	1.5
Contrast, Contrastmono	Contrast	59	**18**	65	2.4	3.0	5.7	4.3	1.1	1.3	**2.5**	1.9
Elaboration:-additional, -definition, -example, -gen-spec,-obj-attribute	Elaboration	62	**155**	645	2.6	3.8	5.7	4.0	.9	.9	**1.8**	1.6
Enablement	Enablement	0	**1**	11			2.0	2.3			**1.0**	1.3
Evaluation: -inferred, -subjective	Evaluation	25	**0**	0	2.9				1.4			
Explanation: -evidence, -reason, -argument	Explanation	55	**3**	10	2.6		5.3	4.8	1.1		**1.3**	1.0
Joint	Joint	12	**35**	151	3.2		5.6	4.1	.6		**1.6**	1.5
Manner-means	Manner-means	8	**3**	9	2.4		4.7	5.1	1.0		**1.3**	2.8
Same-unit	Same-unit	0	**5**	62			3.2	4.5			**2.2**	2.7
Span	Span	2	**0**	0	4.0				.3			
Summary	Summary	4	**0**	0	2.0				1.0			
Temporal	Temporal	2	**2**	9	3.5		5.0	3.6	.4		**3.3**	1.8
Topic-comment: -rhetq, -qa, -problem-solution	Topic-Comment	14	**0**	0	2.5				.8			
All relations		291	**248**	1150	2.6	3.4	5.4	3.9	1.00	1.10	**1.73**	1.61

Table 1 Comparison of sub-trees extracted from the CODA parallel corpus and HILDA's discourse analysis

Evaluation

Evaluation of the comparable sub-trees is listed in Table 2. Relations that have not been listed have a precision and recall of 0. The R-column describes the comparable sub-trees with matching discourse relation; and the RT-column describes the comparable sub-trees with matching discourse relation and EDUs found in their left and right children.

Overall performance is rather poor: precision significantly drops between R and RT matching sub-trees, which indicates that the discourse structure differs. The difference in discourse structure is also indicated by the 84 comparable sub-trees with different EDUs in their left and right children, when EDUs differ, the discourse relation is more likely to differ as well.

HILDA's RST trees contained about 4 times more sub-trees than CODA's RST trees. CODA's EDUs are exactly one segment, whereas HILDA's EDUs are parts of segments. Therefor CODA's manual annotation describes the medium- and high-level discourse, whereas HILDA's analysis describes the low-, medium- and high-level discourse. I.e. HILDA describes the discourse relation of parts of sentences, sentences and segments; whereas CODA manual annotation describes the discourse relation of segments.

	Precision			Recall		
Relation	R	RT	**RTO**	R	RT	**RTO**
Attribution	.20	.20	**.10**	.17	.17	**.08**
Condition	1.00	.80	**.80**	.21	.17	**.17**
Contrast	.39	.22	**.11**	.12	.07	**.03**
Elaboration	.23	.08	**.08**	.58	.21	**.21**
Explanation	.33			.02		
Joint	.09			.25		
Total	.22	.09	**.08**	.19	.08	**.07**

Table 2 Evaluation of HILDA's discourse analysis

The comparison of Table 1 indicates that RTO sub-trees are on average shallower than C sub-trees. Thus differences occur more often when sub-trees are deeper; in effect those are sub-trees describing the high-level discourse, which may indicate that HILDA's high-level discourse analysis is weaker.

CODA		HILDA	
S0	The sense in his becoming brave if he is to get no credit for it involves an important detail of man' s make	*S00*	The sense in his becoming brave if he is to get no credit
		S01	for it involves an important detail of man ' s make
S1	which we have not yet touched upon:	*S1*	which we have not yet touched upon:
{Elaboration *S0*N *S1*}		{Elaboration *S00*N {Elaboration *S01 S1*}}	

Table 3 Structural differences between CODA and HILDA

An example of the effect of more fine grained EDUs is listed in Table 3 based on the paragraph "31-34(Twain-part1_2)" (Stoyanchev and Piwek 2010b). HILDA segmented CODA's left child's EDU into two separate EDUs, introducing a difference in the discourse structure.

CODA's sub-trees children are of equally depth, whereas HILDA's sub-trees are on average deeper in the left child. For example sub-trees with the Background and Joint relations their balance differs significantly, which may explain the low performance of those relations.

Conclusions

The extracted a discourse corpus form the CODA parallel corpus provides an extra source for training and evaluating automatic discourse analysers. HILDA's performance on this corpus is currently weak, mainly due to structural differences. We identified the nature of those differences: the tree construction needs to be reconsidered as the discourse structure has an important impact on the overall performance of HILDA.

Future work

Improve HILDA's segmentation, such that the remaining 36 monologue texts can be parsed. Use the CODA parallel corpus to train HILDA's classifiers and revise HILDA's tree-construction algorithm, such that: 1) sub-trees do not span the right-hand side of one segment and the left-hand side of another segment, in order to prevent situations such as described in Table 3; 2) shift the focus from low-level analysis to high-level analysis; and 3) introduce guidance for generation of more balanced RST trees and for prioritisation of discourse relations. The introduction of rule-based tree-construction without leveraging the computational complexity (LeThanh, Abeysinghe, and Huyck 2004), combined with the information from the classifiers may improve HILDA's performance.

Acknowledgements

We would like to thank Dr. Stoyanchev for her visit in February 2011 to the NII, where she introduced the current state of CODA and pointed out possible future research that lay basis on this case study.

References

Rus, V. et al. 2010. The first question generation shared task evaluation challenge. In *Proceedings of the Sixth International Natural Language Generation Conference,* 251-257. Trim Castle, Ireland.

Rus, V. and Graesser, A.C. 2009. The Question Generation Task and Evaluation Challenge, Workshop Report, ISBN: 978-0-615-27428-7, Institute for Intelligent Systems, Memphis, TN.

Heilman, M. 2011. Automatic Factual Question Generation from Text. Ph.D. diss., Language Technologies Institute, School of Computer Science, Carnegie Mellon University, Pittsburgh, PA.

Piwek, P. and Stoyanchev, S. 2010a, Question generation in the CODA project. In *Proceedings of QG2010: The Third Workshop on Question Generation,* 29-34. Pitsburgh, PA.

Mannem, P., Prasad, R.P. and Joshi, A. 2010. Question Generation from Paragraphs at UPenn. In *Proceedings of QG2010: The Third Workshop on Question Generation,* 84-91. Pitsburgh, PA.

Soricut, R. and Marcu, D. 2003. Sentence level discourse parsing using syntactic and lexical information. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, 149-156. Stroudsburg, PA.

Reitter, D. 2003. Simple Signals for Complex Rhetorics: On Rhetorical Analysis with Rich-Feature Support Vector Models. *LDV-Forum, GLDV-Journal for Computational Linguistics and Language Technology* 18(1/2): 38-52

LeThanh, H., Abeysinghe, G. and Huyck, C. 2004. Generating discourse structures for written texts. In *Proceedings of the 20th international conference on Computational Linguistics,* 329-335. Stroudsburg, PA.

Hernault, H., Prendinger, H., duVerle, D. and Ishizuka, M. 2010. HILDA: A discourse parser using Support Vector Machine classification. *Dialogue and Discourse*, 1(3): 1-33

Hernault, H., Bollegala, D. and Ishizuka, M. 2011. Semi-supervised Discourse Relation Classification with Structural Learning. *Lecture Notes in Computer Science*: Springer, 340-352

Stoyanchev, S. and Piwek, P. 2010b. Constructing the CODA corpus: A parallel corpus of monologues and expository dialogues. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC),* paper-127. Malta