Semantics Representation in a Sentence with Concept Relational Model (CRM)

Rusli Abdullah[2], Jamaliah Abdul Hamid[1], Mohd. Hasan Selamat[2], Hamidah Ibrahim[2] and Ungku Azmi Ungku Chulan[2]

[1]Faculty of Educational Studies
Universiti Putra Malaysia, 43400 UPM Serdang, Selangor
Tel: 03-89468177, Fax: 03-89468246
E-mail: aliah@putra.edu.my

[2]Faculty of Computer Science and Information Technology
Universiti Putra Malaysia, 43400 UPM Serdang, Selangor
Tel: 03-89466555, Fax: 03-89466576
E-mail: rusli@fsktm.upm.edu.my, hasan@fsktm.upm.edu.my, hamidah@fsktm.upm.edu.my, uauc06@yahoo.com

ABSTRACT

The current way of representing semantics or meaning in a sentence is by using the conceptual graphs. Conceptual graphs define concepts and conceptual relations loosely. This causes ambiguity because a word can be classified as a concept or relation. Ambiguity disrupts the process of recognizing graphs similarity, rendering difficulty to multiple graphs interaction. Relational flow is also altered in conceptual graphs when additional linguistic information is input. Inconsistency of relational flow is caused by the bipartite structure of conceptual graphs that only allows the representation of connection between concept and relations but never between relations per se. To overcome the problem of ambiguity, the concept relational model (CRM) described in this article strictly organizes word classes into three main categories; concept, relation and attribute. To do so, CRM begins by tagging the words in text and proceeds by classifying them according to a predefined mapping. In addition, CRM maintains the consistency of the relational flow by allowing connection between multiple relations as well. CRM then uses a set of canonical graphs to be worked on these newly classified components for the representation of semantics. The overall result is better accuracy in text engineering related task like relation extraction.

Keywords

Conceptual graph, concept relational model, language models, semantic network, semantic representation, Natural Language Processing

1.0 INTRODUCTION

In Natural Language Processing (NLP), a language model is crucial in providing a medium between natural language and computational models. Several language models have been devised to represent the semantics of text. Semantics of text refers to the meaning(s) embedded in the sentences within the text. Statistical language models like n-grams are quite effective in natural text processing because of its basic focus on statistical occurrence of word relations in a text. Statistical language models are not hindered by the structure of language, but unfortunately they can be quite restricted in the interpretation of semantics because they cannot handle complex relationships.

Non statistical language model on the other hand, relies on the structure of language to succeed. It works by modeling the representation of meaning within text (semantics) via the manipulation of symbolic meaning captured in the relationship between principal and functional concepts in the text. One of the main challenges of developing a non-statistical language model is deciding what each symbol represents and how these symbols interact in the formation of semantics.

2.0 SEMANTIC REPRESENTATION

The widely used language models are the semantic network (Brachman, 1977) and conceptual graphs (Sowa, 2000; Sowa, 1992; Sowa, 1984). Due to its versatility, conceptual graphs have been employed in many applications related to text processing. This includes relation extraction, text mining (Montes-y-Gomez, Gelbukh & Lopez-Lopez, 2002) and semantic parsing (Sowa & Way, 1986). Concepts in conceptual graph are loosely defined (Sowa, 1984). As such, a concept can either be a noun, verb or adjective. This can result to a variety of ways when representing the same semantics of text. For instance, in the attempt of modeling the phrase 'The pin is blue' (Figure 1).

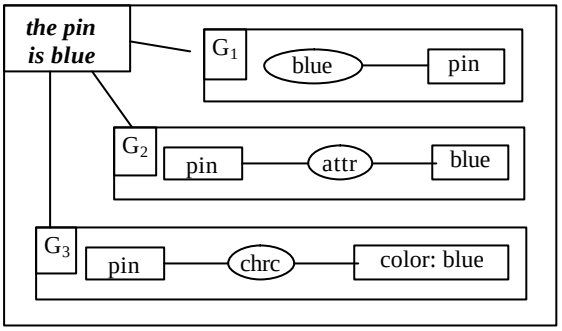


Figure 1: Different Structures of Similar Semantics

By allowing this freedom in denoting concepts, consistency is sacrificed. This leads to difficulty in determining whether graphs of different structures share the same semantics (Montes-y-Gomez, Gelbukh & Lopez-Lopez, 2001). In figure 2, both graphs G_1 and G_3 have the same meaning, but no overlapping structures transpire. 'blue' in G_1 is a conceptual relation, while in G_3, it is a concept. As a result, these two graphs are considered different when they are in fact semantically the same.

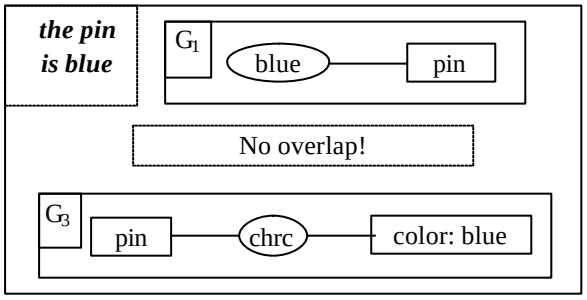


Figure 2: Finding Graph Similarity

3.0 CONCEPT RELATIONAL MODEL

The immediate problem was to develop a non statistical NLP model that provides consistency of representation for the semantics of concepts based on the relationships. This gave rise to Concept Relational Model (CRM). CRM is devised in the effort to introduce simplicity and consistency to language modeling. CRM is made of three components: concept, relation and attribute (Figure 3). CRM only regards noun phrases as concepts (Reinberger, Spyns & Pretorious, 2004; Zhou & Chu, 2003).

Relations imply the connection between concepts.

Example:

⟨Amy ate apples⟩ is modeled as ⟨concept, relation, and concept⟩.

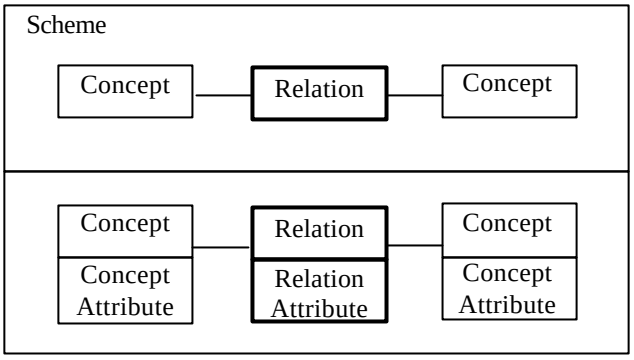


Figure 3: Concept and Relation Attribute

CRM treats the elements of text that are perceived as relations as connectors. The notion of 'connectors' have been used by other researchers as well. Connectors are made of verb (Girju & Moldovan, 2002), preposition (Roberts, 2005; Berland & Charniak, 1999), conjunction (Hearst, 1992), certain types of pronoun (Siddhartan, 2002), comma (Hearst, 1992) and apostrophe (Berland & Charniak, 1999). Attribute can be of two types: concept attribute and relation attribute. Concept attribute modifies the semantics of a concept. Below (Figure 4), the concept 'apple' is modified by '10' and 'sweet'. Therefore, '10' and 'sweet' are both concept attributes. The concept is 'apples'. See Figure 4.

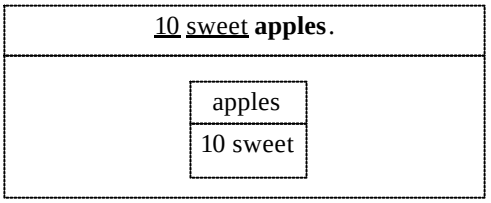


Figure 4: Concept Attribute

Contrary to concept attribute, relation attribute modifies the meaning of a relation. For example in Figure 5, 'hungrily' modifies the relation 'eat'.

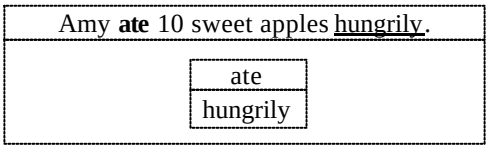


Figure 5: Relation Attribute

An attribute contained within a concept or relation can be subsumed. As such, two sentences, although quite different, but still share similar concepts and relations are regarded to be 'generally' the same. The illustration demonstrates this idea (Figure 6). Both sentences have the same set of concepts and relation. As such, by allowing the subsumption of attributes in CRM, simplicity may be achieved.

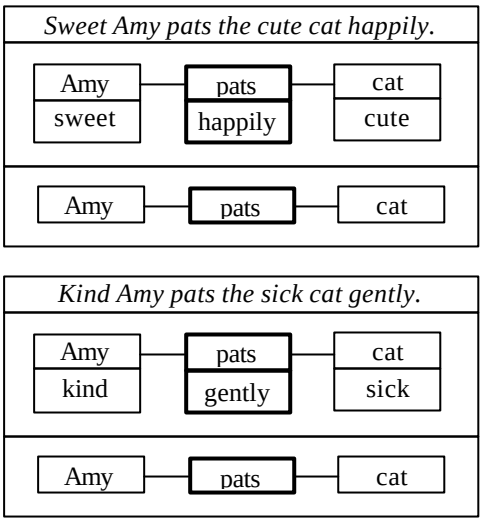


Figure 6: Similar Sentences

4.0 APPLYING TAG SET IN THE CRM

The usual part-of-speech (pos) tags categorise words into nouns, verbs, etc. CRM on the other hand divides word classes into concept (C), relation (R), and attributes (A_c for Attribute of Concept; and A_R for Attibute of Relation). The division is achieved by classifying the part of speech tags into the following concept relational model tag-set or CRM-Tag:

POS-Tag	CRM-Tag
NN|NNP|NNPS|NNS	C
VB|VBD|VBG|VBN|VBP|VBZ	R
JJ | JJR | JJS	A_C
RB | RBR | RBS	A_R
PRP | PRP$	C
CC	R
IN	R
CD	A_C
POS	R
TO	R
WDT|WP|WP$ | WRB	R
RP	R

In CRM, word classes like determiner (DT) and interjection (UH) is omitted since they are regarded to be trivial in term of content (Hearst, 1992). In the illustration (Figure 7), the tags for words in the sentence are converted from the common pos tag set into the CRM-tag set.

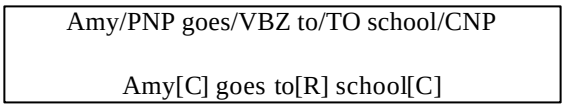


Figure 7: Conversion of pos-tags to CRM-tags

By classifying words in this manner, semantics in CRM may be represented in its most consistent form.

5.0 CANONICAL GRAPHS IN CRM

Canonical graphs define the allowed structural arrangement of concepts and relations. It identifies deviant structures from those acceptable ones, and by this virtue, minimizes erroneous meaning representation in text processing.

Inspired by the idea of canonical graph (Sowa, 1984), a set of canonical graph or structures are defined by CRM to initiate its probable usage in NLP. The set of graphs are shown in Figures 8.1 to 8.6. Each depicts acceptable canonical relationship between concepts, relations, and attributes.

1. Intransitive Verb: R_1 (Figure 8.1)

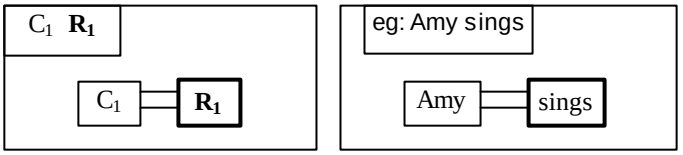


Figure 8.1: Intransitive Verb

2. Transitive Verb: R_1 (Figure 8.2)

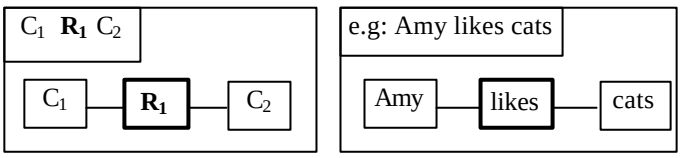
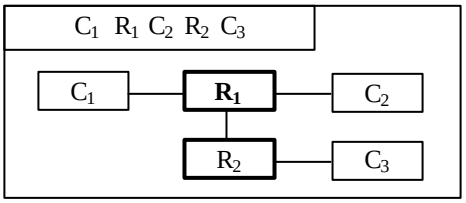


Figure 8.2: Transitive Verb

3. Ditransitive Verb: R_1 (Figure 8.3)



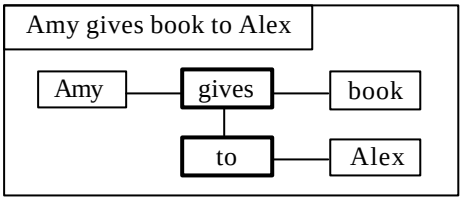


Figure 8.3: Ditransitive Verb

4. Adverbial Attachment: R_2 (Figure 8.4)

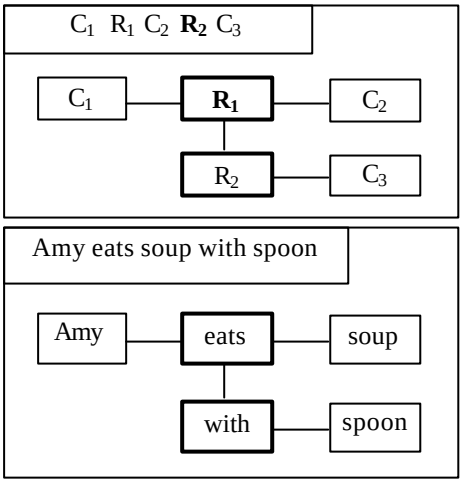


Figure 8.4: Adverbial Attachment

5. Adjectival Attachment: R_2 (Figure 8.5)

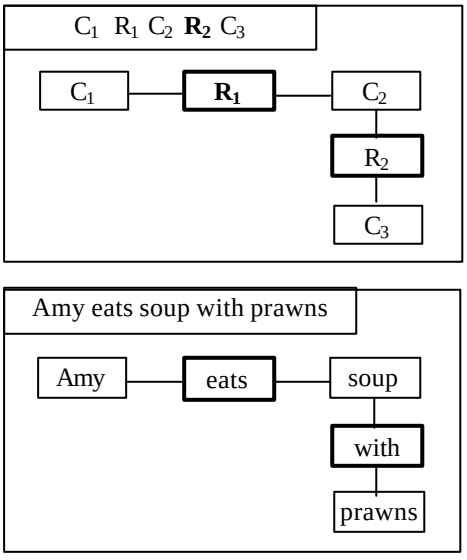
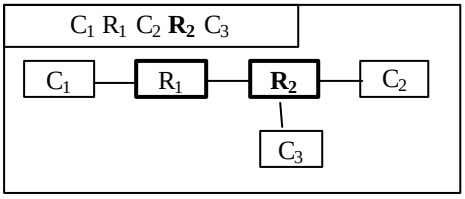


Figure 8.5: Adjectival Attachment

6. Conjunction: R_{N-1} (Figure 8.6)



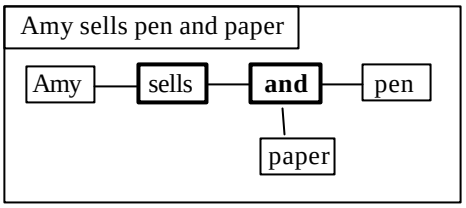


Figure 8.6: Conjunction

6.0 IMPROVING THE CONSISTENCY OF CONCEPTUAL GRAPHS

While the canonical graphs set delimiters to the number of acceptable relationships between concepts, relations, and their attributes, they do not however point to the direction of the flow between those relationships. Directional flow is important among other things to tell us the sequence of the relationships, especially when there are more than two or three concepts.

To note, the flow in conceptual graphs might change when additional information is appended to the original graphs. This can be seen in the illustration (Figure 9). The second conceptual graph (G_2) is derived from the first one (G_1) by adding some information. Apparently, the flow between the two concepts 'Amy' and 'poem' change when semantics is extended.

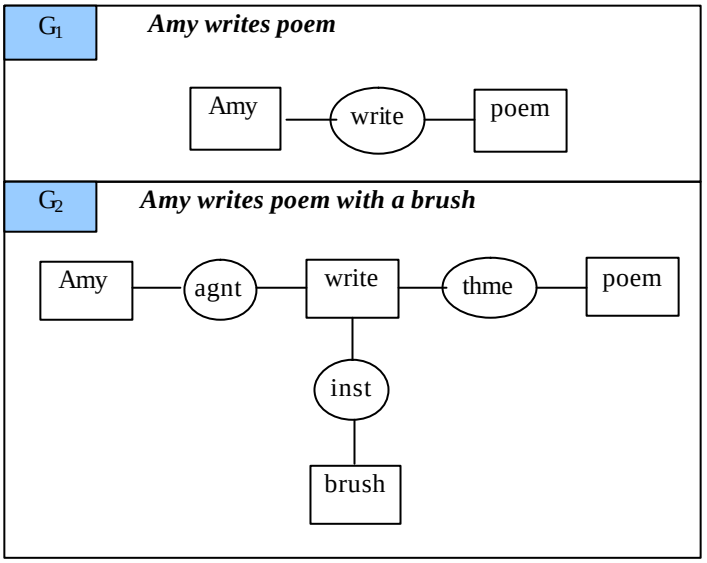


Figure 9: Change of Flow

The reason of this comes from the fact that conceptual graph is innately a 'bipartite graph'. Link between nodes of the same type is not allowed. Thus, a link between two relations is prohibited in conceptual graph (that 'write a poem' and 'write with a brush'). Changing the flow risks the possibility of erroneous interpretation whenever the graph is modified. As an alternative, CRM uses the link between relations to represent semantics. This way, the flow can be maintained without risking inconsistency (Figure 10).

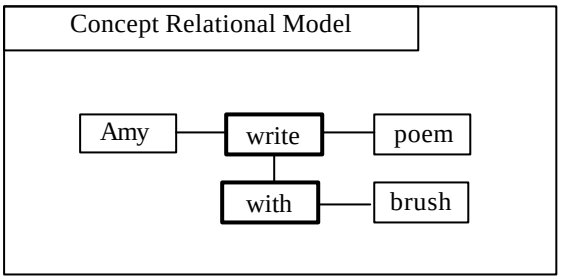


Figure 10: Alternative to Maintain Flow

7.0 CONCLUSION

The concept relational model (CRM) offers a language model that organizes the information in text into three categories; concept, relation and attribute. The classification is done via the part-of-speech tags of words. Seven kinds of canonical graph are generated for the concept relational model involving the use of verbs and conjunctions are proposed. Although it is far from comprehensive, it can act as a guide for the development of other canonical graphs. The graph assumes that a sentence S is represented using the concept relational model, whereby $S = C_1 R_1 \ldots R_{N-1} C_N$. At the moment the CRM is limited to English only. For that, the model is more compact but not as robust as the conceptual graphs. However, CRM overcomes the ambiguity and inconsistency of conceptual graphs. By doing so, better accuracy can be achieved in text engineering related task like relation extraction. This leads to better measurement of similarity for similar graphs. As such, the process of integrating similar graphs is enhanced**.**

8.0 REFERENCES

Berland, R. & Charniak, E. C. (1999). Finding Parts in Very Large Corpora. *In Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics, ACL 1999.*

Brachman, R. (1977). What's a Concept: Structural Foundations for Semantic Networks. *International Journal of Man-Machine Studies, 9*.

Hearst, M. A. (1992). Automatic Acquisition of Hyponyms from Large Text Corpora. In *Proceedings of the Fourteenth International Conference on Computational Linguistics, pages 539—545*. Nantes, France, July 1992.

Montes-y-Gomez, M., Gelbukh, A., Lopez-Lopez, A. & Baeza-Yates, R. (2001). Flexible Comparison of Conceptual Graphs. *Proceeding of DEXA – 2001, 12th International Conference and Workshop on Database and Expert System Applications*, Munich, Germany, September 2001.

Montes-y-Gomez, M., Gelbukh, A. & Lopez-Lopez, A. (2002). Text Mining at Detail Level using Conceptual Graphs. *10th International Conference on Conceptual Structures*, ICCS 2002 Borovets, Bulgaria, July 15-19, 2002.

Reinberger, M. L., Spyns, P. & Pretorious, A. J. (2004). Automatic Initiation of an Ontology. *In Proceedings of ODBase* 2004.

Roberts, A. (2005). Learning Parts and Wholes from Biomedical Texts. *In Proceedings of the 8th Research Colloquium of the UK special-interest group in Computational Linguistics (CLUK-05)*, pages 63-70, Manchester, UK, January 2005. CLUK.

Siddharthan, A. (2002). An architecture for a text simplification system. In *Proceedings of the Language Engineering Conference 2002 (LEC 2002)*, pages 64-71.

Sowa, J. F. (1984). *Conceptual structures: Information processing in mind and machine.* Addison Wesley.

Sowa, J. F. (1992). Conceptual Graphs Summary. In P. Eklund, T. Nagle, J. Nagle, and L. Gerholz, (eds.) *Conceptual structures: Current research and practice*. Ellis Horwood, 1992, pages 3-52.

Sowa, J. F. (2000). *Knowledge Representation: Logical, Philosophical and Computational Foundations.* Brooks/ Cole. Thomson Learning.

Sowa, J. F. & Way, E. C. (1986). Implementing a semantic interpreter using conceptual graphs. *IBM Journal of Research and Development 30:1*, January, 1986.

Zou, Q. & Chu W. (2003). IndexFinder: A knowledge-based method for indexing clinical texts. *American Medical International Association (AMIA) 2003*.