SEMANTICS REPRESENTATION IN A SENTENCE WITH CONCEPT RELATIONAL MODEL (CRM)

Rusli Abdullah, Mohd Hasan Selamat, Hamidah Ibrahim, Ungku Azmi Ungku Chulan and Nurul Amelina Nasharuddin

Faculty of Computer Science and Information Technology
Universiti Putra Malaysia

rusli@fsktm.upm.edu.my
hasan@fsktm.upm.edu.my
hamidah@fsktm.upm.edu.my
uauc06@yahoo.com
nurulamelina@gmail.com

Jamaliah Abdul Hamid

Faculty of Educational Studies
Universiti Putra Malaysia

aliah@putra.edu.my

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.

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.

One of the non statistical language models for representing semantics is by using the conceptual graphs. But conceptual graphs define concepts and conceptual relations loosely. This will creates ambiguity in classifying a word as either a concept or relation. The model maintains the consistency of the relational flow by allowing connections between multiple relations such as C-R-R; R-C-R; C-R-R-C. Ordinary graphs normally disallow multiple relations since relations are treated linearly as in C-R-C. We proposed the use of Concept Relational Model (CRM) to overcome the problem of ambiguity NLP.

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), semantic parsing (Sowa & Way, 1986) and graph representation which used to model a situation of information

or knowledge relationship (Chen & Mugnier, 2008). 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' (See Figure 1).

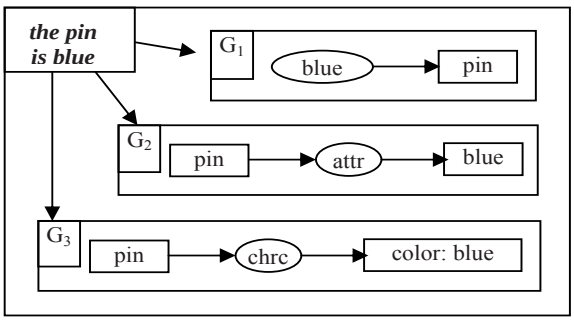

Fig. 1. Different Structure of Similar Semantics

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 & Bueza-Yates 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.

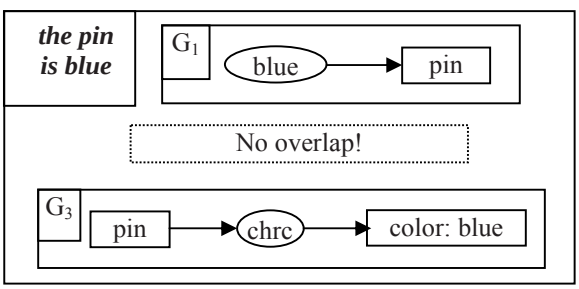

Fig. 2. Finding Graph Similarities

CONCEPT RELATIONAL MODEL AND ITS RELATIONSHIPS

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 (See 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, concept⟩. The scheme of the CRM attributes for the example is also shown in Figure 3, 4, 5 and 6.

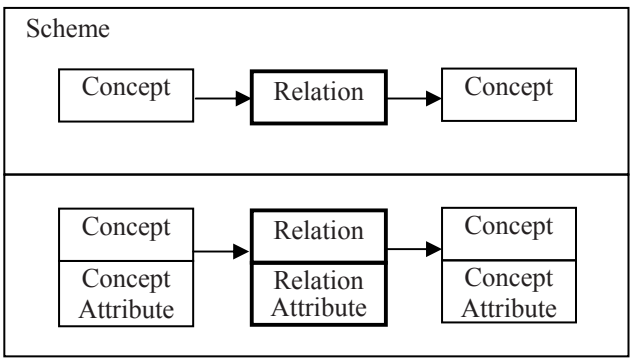

Fig. 3. Concept and Relation Attribute

CRM treats the elements of text that are perceived as relations and 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. Figure 4 below shows the concept 'apple' is modified by '10' and 'sweet'. Therefore, '10' and 'sweet' are both concept attributes. The concept is 'apples'.

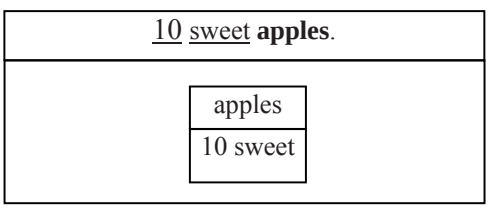

Fig. 4. Concept Attribute

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

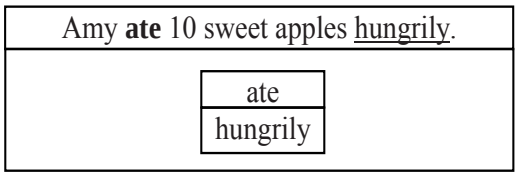

Fig. 5. Relation Attribute

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 (See 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.

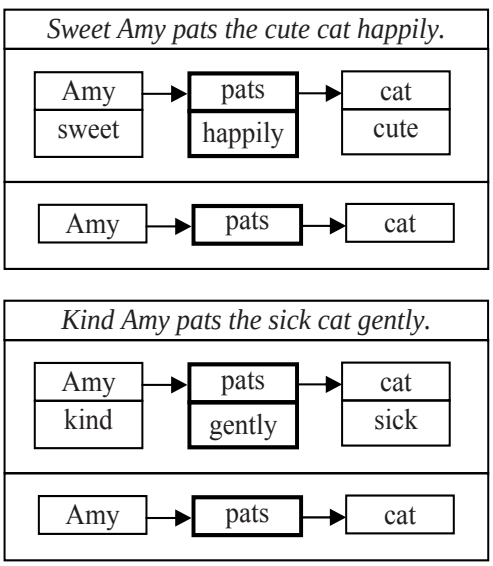

Fig. 6. Difference Sentences with a Same CRM

APPLYING TAG SET IN THE CRM

The usual part-of-speech (POS) tags categorize 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 Attribute of Relation). The division is achieved by classifying the part of speech tags into the following concept relational model tag-set or CRM-Tag:

Table 1. Part-of-speech (POS) Tags

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.

POS-Tag: Amy/PNP goes/VBZ to/TO school/CNP

CRM-Tag: Amy[C] goes to[R] school[C]

Fig. 7. Conversions of Pos-Tags to CRM-Tags

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

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 to 13. Each depicts acceptable canonical relationship between concepts, relations, and attributes.

1. Intransitive Verb: R_1

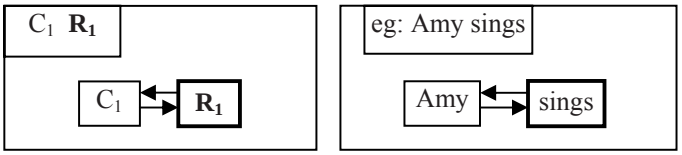

Fig. 8. Intransitive Verb

2. Transitive Verb: R_1

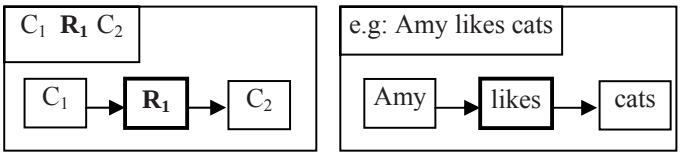

Fig. 9. Transitive Verb

3. Ditransitive Verb: R_1

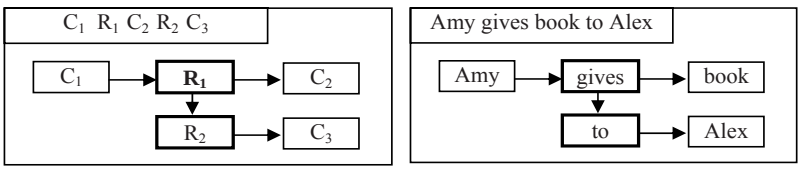

Fig. 10. Ditransitive Verb

4. Adverbial Attachment: R_2

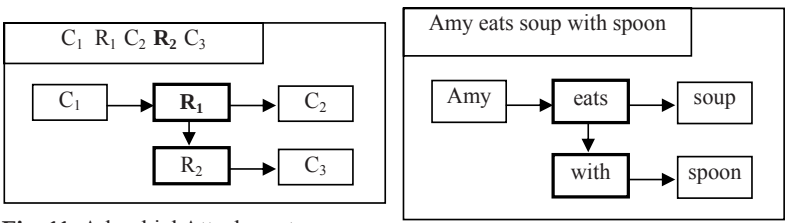

Fig. 11. Adverbial Attachment

5. Adjectival Attachment: R_2

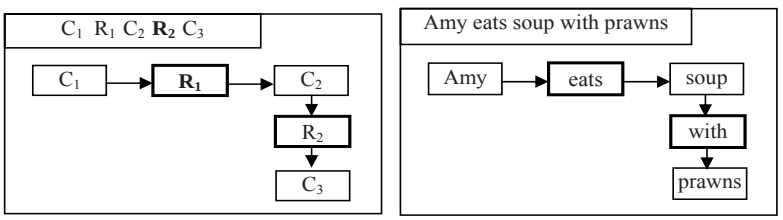

Fig. 12. Adjectival Attachment

6. Conjunction: R_{N-1}

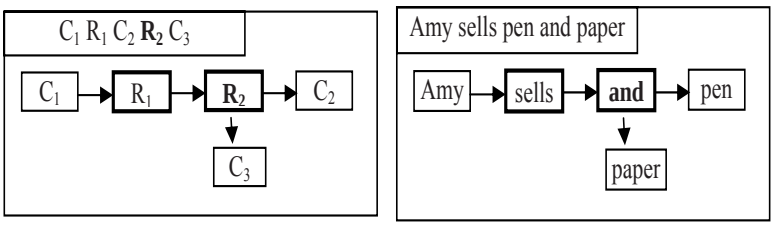

Fig. 13. Conjunction

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 Fig. 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.

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').

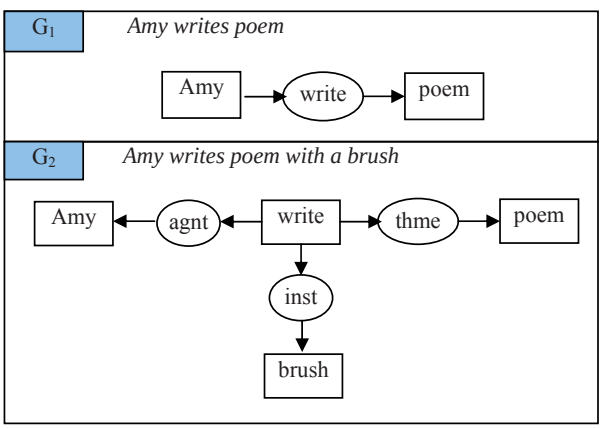

Fig. 9. Change of Flow

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 (Fig. 10).

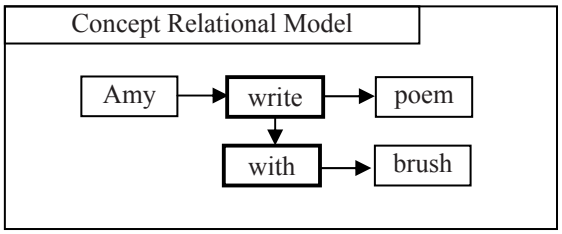

Fig. 10. Alternative to Maintain Flow

CONCLUSION

Traditional conceptual graphs define concepts and conceptual relations loosely. This causes ambiguity because a word can be classified as a concept or relation. This ambiguity further complicates analysis of relations between concepts, if concepts themselves are posited as relations. The proposed Concept Relational Model (CRM) reduces ambiguity by strictly organizing word classes into three main categories; concept, relation and attribute. The classification is done via the part-of-speech tagging of words in texts and proceeds by classifying them according to a predefined mapping. Six kinds of canonical graph which are generated for the CRM involving the use of verbs and conjunctions are proposed. The model then proposes a set of canonical graphs to be used on these newly classified components for the representation of semantics.

Although it is far from comprehensive, it can act as a guide for the development of other canonical graphs. 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**.**

REFERENCES

Berland, M., & Charniak, E. (1999). Finding parts in very large corpora. In *Proceedings of the 37th Annual Meeting of the Association For Computational Linguistics on Computational Linguistics*. Annual Meeting of the ACL. Association for Computational Linguistics, Morristown, NJ, (pp. 57-64).

Brachman, R. (1977). What's a concept: Structural foundations for semantic networks. *International Journal of Man-Machine Studies, 9*(2), 157-152.

Chein, M., & Mugnier, M. L. (2008). *Graph-based knowledge representation: Computational foundations of conceptual graphs (Advanced Information and Knowledge Processing)*. Springer-Verlag.

Girju, R., & Moldovan, D.I. (2002). Text mining for causal relations. In *Proceedings of the Fifteenth International Research Society Conference (FLAIR)*, (pp. 360-364). Florida, USA.

Hearst, M.A. (1992). Automatic acquisition of hyponyms from large text corpora. In *Proceedings of the Fourteenth International Conference on Computational Linguistics,* (pp. 539 - 545). Nantes, France.

M. Montes y Gómez, A. Gelbukh, A. López López, Ricardo Baeza-Yates. Flexible comparison of conceptual graphs. In: Mayr, H.C., Lazansky, J., Quirchmayr, G., Vogel, P. (Eds.), Proc. DEXA-2001, 12th International Conference and Workshop on Database and Expert Systems Applications. *Lecture Notes in Computer Science*, N 2113, Springer-Verlag, (pp. 102-111).

Montes-y-Gomez, M., Gelbukh, A., & Lopez-Lopez, A. (2002). Text mining at detail level using conceptual graphs. In *10th International Conference on Conceptual Structures,* ICCS 2002. Borovets, Bulgaria.

Reinberger, M.-L., Spyns, P., Pretorius, A., & Daelemans, W. (2004). Automatic initiation of an ontology. In R. Meersman et al. (Eds.), *On the move to Meaningful Internet Systems 2004: CooPIS, DOA and ODBASE (Part I)*, LNCS (Vol. 3290, pp. 600-617). Springer-Verlag.

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)*, (pp. 63-70). Manchester, UK.

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

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*, (pp. 3-52). Ellis Horwood.

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), 57.

Zou, Q., & Chu, W. (2003). IndexFinder: A knowledge-based method for indexing clinical texts. In *Proceedings of American Medical International Association (AMIA) Annual Symposium 2003*, (pp. 763-767). Washington D.C, USA.