Extracting Concrete Senses of Lexicon through Measurement of Conceptual Similarity in Ontologies

Siaw-Fong Chung[*], **Laurent Prévot**[**], **Mingwei Xu**[#], **Kathleen Ahrens**[*], **Shu-Kai Hsieh**[##] **and Chu-Ren Huang**[#]

[*]National Taiwan University, No.1, Sec. 4, Roosevelt Road, Taipei, Taiwan 106.
[**]CLLE-ERSS, CNRS, Toulouse, France.
[#]Institute of Linguistics, Academia Sinica, 128 Academia Road, Sec. 2, Nankang, Taipei, Taiwan 115.
[##]Department of English, National 162, He-ping East Road, Sec. 1, Taipei, Taiwan 10610.
E-mail: siawfongchung@gmail.com, prevot@univ-tlse2.fr, javan@gate.sinica.edu.tw, kathleenahrens@yahoo.com, shukai@gmail.com, churen@sinica.edu.tw

Abstract

The measurement of conceptual similarity in a hierarchical structure has been proposed by studies such as Wu and Palmer (1994) which have been summarized and evaluated in Budanisky and Hirst (2006). The present study applies the measurement of conceptual similarity to conceptual metaphor research by comparing concreteness of ontological resource nodes to several prototypical concrete nodes selected by human subjects. Here, the purpose of comparing conceptual similarity between nodes is to select a concrete sense for a word which is used metaphorically. Through using WordNet-SUMO interface such as SinicaBow (Huang, Chang and Lee, 2004), concrete senses of a lexicon will be selected once its SUMO nodes have been compared in terms of conceptual similarity with the prototypical concrete nodes. This study has strong implications for the interaction of psycholinguistic and computational linguistic fields in conceptual metaphor research.

1. Introduction

The Conceptual Metaphor Theory (Lakoff, 1993; Lakoff and Johnson, 1980) proposes that source domains are usually more concrete than target domains. According to this view, a source domain mapping in a metaphorical expression should have two meanings: an abstract meaning from the target domain and a concrete meaning from the source domain. If this is the case, the different senses of a lexicon with metaphorical reading should in fact comprise both concrete and abstract meanings. For example, a Chinese metaphorical expression such as *cheng2zhang3* 'grow/growth' (in *jing1ji4 cheng2zhang3* 'the economy grows') has the following WordNet meanings (Fellbaum, 1998) in (1). Their corresponding ontological nodes (taken from SUMO or Suggested Upper Merged Ontology, Niles and Pease, 2001) are given in brackets.[1]

(1) Meanings of *cheng2zhang3* 'grow/growth' provided by WordNet.

  (a) become bigger or greater in amount (SUMO= 'Increasing')

  (b) grow old or older (SUMO= 'Increasing')

  (c) develop and reach maturity; undergo maturation (SUMO= 'Growth')

  (d) grow emotionally or mature (SUMO= 'Growth')

  (e) a change resulting in an increase (SUMO= 'Growth')

  (f) the process of an individual organism growing organically; a purely biological unfolding of events involved in an organism changing gradually from a simple to a more complex level (SUMO= 'IntentionalPsychologicalProcess')

Among these senses, we can decide intuitively some senses are possibly literal (b, c, d) which some others are possibly metaphorical (a, e). Some senses (such as the last sense of (f)) cannot be decisively determined, falling between a literal and a metaphorical meaning, because this sense is related to 'emotion' or 'reasoning' in the brain, neither of which refer to the actual concept of growing in size.

The Conceptual Metaphor Theory does not provide clear criteria for how concreteness of a source domain can be should be evaluated. Chung (2007) suggests that the concrete senses of a word carry the source domain information if this word is used metaphorically. In this

---

[1] Since all meanings in (1) are meanings provided by the English WordNet for the Chinese lexicon of *cheng2zhang3* 'grow/growth' through the interface provided by SinicaBow (http://bow.sinica.edu.tw), in the future, we hope to use the definitions from the Chinese WordNet (available at http://cwn.ling.sinica.edu.tw/). WordNet is available at http://wordnet.princeton.edu/ while SUMO is available at http://www.ontologyportal.org/.

study, we further suggest that the concrete source domain of a metaphorical expression can be uncovered by using concrete sense disambiguation such that proposed in example (1) above. However, since measurement of concreteness has not been carried out in previous study and there is no attempt to determine concreteness automatically, we outline a criteria-based method to specify concreteness measure, as it is an important issue in conceptual metaphor theory (Lakoff, 1993; Lakoff and Johnson, 1980). In this paper, we suggest a way to determine concrete sense from among the many senses of a word.

2. Formulas and Methodology for Concreteness Measure

Budanisky and Hirst (2006:19) provided the formula in Figure 1 below so that conceptual similarity and conceptual distance between $C1$ and $C2$ can be calculated.[2]

$$(2a) \quad ConSim\ (C1, C2) = \frac{2 * N3}{N1 + N2 + 2 * N3}$$

$$(2b) \quad Distance(C1, C2) = 1 - ConSim\ (C1, C2)$$

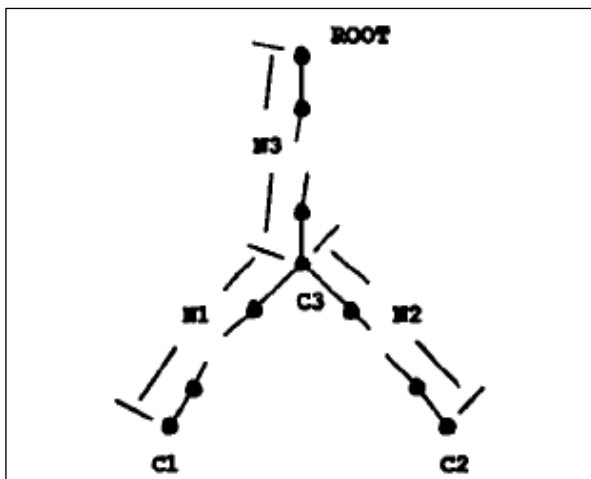
Figure 1: Figure used in Wu and Palmer (1994: 136) for Measurement of Conceptual Similarity

In Figure 1, there are four nodes – Root, $C1$, $C2$ and $C3$. These nodes have distances marked with $N1$, $N2$ and $N3$ respectively (which may comprise several nodes in one measure of distance). The formulas used by Budanisky and Hirst (2006:19) are given in (2) where ConSim is an abbreviation of "Conceptual Similarity." This paper will follow these formulas in (2). However, since the formulas in (2) require the calculation of conceptual similarity before the calculation of the conceptual distance, the reverse information is more suitable for our purpose, if we have our own initial definition of distance. Our definition of distance is a simple calculation of the number of nodes to the root of 'Entity.' For example, in Figure 2 below, the distance between $C1$ and $C3$ ($D1$) is 3 because $C1$ is the third node from $C3$.

---

[2] The formula by Budanisky and Hirst, (2006:19) was a modified version from Wu and Palmer (1994: 136).

The distance between $C1$ and $C2$ is the sum of $D1$ and $D2$. This sum is further divided by $D3$, which is the distance to the root of the hierarchy.[3] The formulas are given in (3) below. (3a) shows the calculation of 'conceptual distance' which is based on our own definition of 'distance.' (3b) shows the calculation of 'conceptual similarity,' which is a transformation of the formula in (2b).[4]

$$(3a) \quad ConceptualDistance(C1, C2) = \frac{D1 + D2}{Exp\ (D3)}$$

$$(3b) \quad Conceptual\ Similarity\ (C1, C2) = \frac{1}{1 + Dis \tan ce}$$

The purpose of comparing conceptual similarity between different nodes, it must be stressed, is to select one concrete sense for a word which is used metaphorically. This concrete sense is believed to carry the original meanings of the metaphor. Therefore, the study proposed herein can be applied to automatic source domain (i.e., the concrete domain)

---

[3] For $D3$, an 'exponential' value has been added, because it transforms the number of $D3$ to a smaller value and this will increase the sum of $D1$ and $D3$ in the calculation process. 'Exponential' is added in order to avoid cases where $D1$ and $D2$ are short, indicating that $C1$ and $C2$ appear at nodes closer to the root. When this happens, calculation without 'exponential' will return a high conceptual similarity score which will create an artificially higher value which is in fact incorrect.

[4] An example of program is shown in below (explanations are given after the symbol of #). The squared node ('InternalChange') is the node at $C3$, where the paths of $D1$ and $D2$ meet. For $C1$ ('Growth'), it is the forth node after 'Process' (thus, $D1=4$). $C2$ ('Cooking') is one of the prototypical concrete nodes selected through human ratings. It is the third node after process (thus, $D2=3$).

```
D1: Entity Physical Process
    InternalChange (C3)
    BiologicalProcess PhysiologicProcess
    AutonomicProcess Growth (C1)
D2: Entity Physical Process
    InternalChange (C3) Creation Making
    Cooking (C2)
D1: 4   D2: 3   D3: 3
Conceptual Distance:
(4+3)/(20.085537)=0.3485095
Conceptual Similarity:
1/(1+0.3485095)=0.74155945
```

If one of the nodes (e.g., 'Cooking') has multi-path, the average for the different paths will first be found.

```
D1: Entity Physical Process (C3)
    InternalChange BiologicalProcess
    PhysiologicProcess AutonomicProcess
    Growth (C1)
D2: Entity Physical Process (C3)
    IntentionalProcess Making Cooking (C2)
D1: 5   D2: 3   D3: 2
Conceptual Distance:
(5+3)/(7.389056)=1.0826823
Conceptual Similarity:
1/(1+1.0826823)=0.48015007
Average =(0.74155945+0.48015007)/2
=0.61085474
```
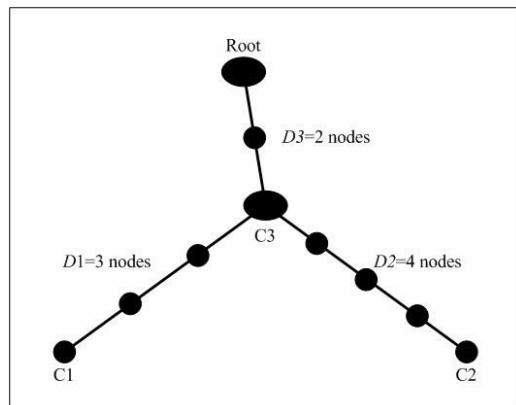
1965

determination in future research.

Figure 2: Measuring Conceptual Similarity between SUMO Nodes

Since also we use a WordNet-SUMO interface provided by SinicaBow (Huang *et al*, 2004), we can measure concreteness of different WordNet senses based on their corresponding SUMO nodes. Before we can implement the formulas in (3) in SUMO, we need to provide a standard delimiting what is concrete. We selected 30 prototypical concrete nodes based on subjects' ratings of the concreteness of the SUMO nodes.

An experiment was run where ten subjects were asked to rate all SUMO nodes (except functions and relations) according to the scale of concreteness from 1 to 7, with 1 being least concrete and 7 being most concrete. A total of 626 SUMO nodes were provided in a single questionnaire to subjects online. From the ratings of the subjects, we selected thirty prototypical concrete nodes which were considered concrete by the subjects. Ten prototypical concrete nodes were selected from 'Abstract' (if they are rated concrete); ten other nodes from 'Object' and the remaining ten from 'Process.' These selected nodes are in average high in their ratings of concreteness. The selected prototypical concrete nodes fall under three main parent nodes, shown in Table 1.

Major Parent Nodes	Prototypical Concrete SUMO Nodes	Immediate Parent Nodes of Prototypical Concrete Nodes (Multi-paths are Separated by '/')	Ratings
Object (*Mean*= 5.15)	Building	SelfConnected Object	5.80
	BodyPart	SelfConnected Object	5.70
	Canine	SelfConnected Object/Agent	5.40
	Educational Organization	Collection/Agent	5.00
	Government	Collection/Agent	4.70
	Organization	Collection/Agent	5.20
	FamilyGroup	Collection/Agent	4.90
	LandArea	Region	5.00
	SaltWaterArea	Region	4.90
	GeographicArea	Region	4.90
Process (*Mean*= 4.98)	Surgery	IntentionalProcess	5.40
	Gesture	IntentionalProcess/Motion	5.20
	Speaking	IntentionalProcess/Motion	4.90
	Selling	IntentionalProcess/DualObjectProcess	4.50
	Cooking	InternalChange/IntentionalProcess	5.00
	Breathing	InternalChange/InternalChange	5.10
	RadiatingLight	Motion	4.70
	Shooting	Motion	5.33
	Covering	Motion	4.90
	Walking	Motion	4.80
Abstract (*Mean*= 4.96)	PrimaryColor	InternalAttribute	4.50
	ColorAttribute	InternalAttribute	4.00
	DiseaseOr Syndrome	InternalAttribute	3.90
	Plan	InternalAttribute	5.00
	Computer Program	Procedure	4.90
	CelsiusDegree	Procedure	5.10
	EuroCent	PhysicalQuantity	5.60
	EuroDollar	PhysicalQuantity	5.60
	Centimeter	PhysicalQuantity	4.90
	UnitedStatesCent	PhysicalQuantity	6.10

Table 1: Selected Prototypical Concrete Nodes and their Average Ratings[5]

---

[5] All nodes are covered except 'Relation' and 'SetorClass' (under 'Abstract') because 'Relation' comprises relational

In Table 1, 'Building,' 'BodyPart' and 'Canine' have the immediate parent node of 'SelfConnectedObject.' 'Educational Organizaion' has two immediate parent nodes ('Collection' and 'Agent'), which means that 'Educational Organization' has two paths (called a 'multi-paths' in Table 1). For prototypical concrete nodes that have more than one path, an average score will be calculated based on these different paths.

As for the measurement of conceptual similarity, each corresponding SUMO node of a WordNet sense will be compared to their respective upper categories of 'Object,' 'Process' and 'Attribute.' For example, one of the WordNet senses for *cheng2zhang3* 'grow/growth' has the corresponding SUMO node of 'Growth.' This SUMO node is under the parent node of 'Process.' Therefore, 'Growth' will be compared to all prototypical concrete nodes under the parent nodes of 'Process' in Table 1, which are 'Surgery,' 'Gesture,' 'Speaking,' 'Selling,' etc. After 'Growth' has been compared to each of the parent nodes of 'Process,' a score will be calculated for each comparison, whereby each score will be obtained by comparing 'Growth' and 'Surgery;' 'Growth' and 'Gesture,' and so on.

'Growth' will therefore have ten scores of conceptual similarity from ten prototypical concrete nodes. Among these ten scores, only the highest will be selected and this highest score will form the highest score for 'Growth' in terms of its concreteness. The higher this number is, the higher the concreteness is. As mentioned, if a prototypical concrete node has multi-paths, the average for the multi-paths will first be computed before finding out the highest score from among the averaged multi-path.

3. Results of Measuring Concreteness of SUMO Nodes

Based on the calculations completed in the previous section, each metaphorical expression (such as *cheng2zhang3* 'grow/growth') will have one final selected concrete sense as defined by the highest score of conceptual similarity when compared to the prototypical concrete nodes. Table 2 below shows the calculation of concreteness for *cheng2zhang3* 'grow/growth,' *qi3fei1* 'take off' and *tan1huan4* 'paralytic.'

	WordNet Definitions	SUMO Nodes (Upper Nodes)	Concrete-ness Scores	Highest Score
cheng2 zhang3 'grow/ growth'	a change resulting in an increase: "the increase is scheduled for next month"	Increasing (Process)	0.77	
	become bigger or greater in amount: "The amount of work increased"	Increasing (Process)	0.77	
	the process of an individual organism growing organically; a purely biological unfolding of events involved in an organism changing gradually from a simple to a more complex level: "he proposed an indicator of osseous development in children"	**Growth (Process)**	**0.98**	**V**
	grow old or older: "She aged gracefully" "we age every day" "what a depressing thought!"	**Growth (Process)**	**0.98**	**V**
	develop and reach maturity; undergo maturation: "He matured fast" "The child grew fast"	**Growth (Process)**	**0.98**	**V**
	grow emotionally or mature: "The child developed beautifully in her new kindergarten" "When he spent a summer at camp, the boy grew noticeably and no longer showed some of his old adolescent behavior"	Intentional Psychological Process (Process)	0.92	
qi3fei1 'take off'	a departure; especially of airplanes:	Motion (Process)	0.91	
	the initial ascent of an airplane as it becomes airborne:	Motion (Process)	0.91	
	depart from the ground, as of an aircraft or balloon: "The plane took off two hours late"	**Transportation (Process)**	**0.93**	**V**
tan1huan4 'paralytic'	a condition marked by uncontrollable tremor:	Pathologic Process (Process)	0.93	
	loss of the ability to move a body part:	**Disease Or Syndrome (Attribute)**	**1.00**	**V**

Table 2: Selected SUMO Nodes with Highest Concreteness Scores from SinicaBow

meanings that are not nodes and 'SetorClass' is not an upper category found among the mappings in the data obtain (thus, including it will cause noises as distance will be increase unnecessarily). These selected SUMO nodes have mean ratings that do not differ from one another, $F_{(2,297)}=0.641$, $p=.527$ (*mean* for 'Object' is 4.96; *mean* for 'Process' is 5.15; and *mean* for 'Abstract' is 4.93). The nodes selected in Table 1, therefore, also cover most of the nodes under 'Object,' 'Process' and 'Abstract.'

From Table 2, the selected concrete SUMO node for *cheng2zhang3* 'grow/growth' is 'Growth.' Three WordNet senses are mapped to this SUMO node (shaded). These nodes are under the category of 'Process' (third column). Since there may be more than one WordNet sense with similar SUMO nodes ('Growth' for example has several senses of *cheng2zhang3* 'grow/growth,' which includes both verbs and nouns), final SUMO nodes will ultimately be selected. These SUMO nodes may correspond to several WordNet senses, meaning that the few WordNet senses under the same SUMO node should have a similar concept of concreteness, which does not contradict the overall results. For *qi3fei1* 'take off' and *tan1huan4* 'paralytic,' the SUMO node selected has only one corresponding WordNet sense. For *qi3fei1* 'take off,' 'Transportation' has the highest score. For *tan1huan4* 'paralytic,' 'DiseaseOrSyndrome' has the highest score.

Based on the results of the calculation, we obtain the senses that are concrete. These senses will help identify the literal meanings of the metaphorical expressions (e.g., 'Growth' for phrase such as *jing1ji4 cheng2zhang3* 'the economy grows;' 'Transportation' for *jing1ji4 qi3fei1* 'the economy takes off;' and 'DiseaseOrSyndrome for *jing1ji4 tan1huan4* 'the economy becomes paralytic').

4.   Conclusion

This paper proposes an innovative way of measuring conceptual similarity between different concepts. Applying this measurement of conceptual similarity to metaphor research will show that cross-disciplinary research can be carried out. In this work, knowledge in psycholinguistics, computational linguistics and metaphor research are combined. Work in this paper will not only contribute to concrete sense disambiguation, it will also contribute to the evaluation of the ontology where distance between nodes will be evaluated in terms of conceptual similarity.

As future work, we hope to propose ways to evaluate the methodology suggested herein. We also propose to extend this analysis of conceptual similarity to WordNet so as to compare whether or not mappings of WordNet and SUMO provided by interface such as SinicaBow (Huang *et al.* 2004) are reliable. If the mappings are consistent, the calculating of their conceptual similarity should also be similar.

5.   Acknowledgements

6.   References

Budanitsky, A. and Hirst, G.. (2006). Evaluating WordNet-based Measures of Semantic Relatedness. *Computational Linguistics*. 32 (1), pp. 13--47.

Chung, S.-F. (2007). *A Corpus-driven Approach to Source Domain Determination*. Ph.D. Dissertation. Graduate Institute of Linguistics, National Taiwan University.

Fellbaum, C. (ed.). (1998). *WordNet: An Electronic Lexical Database*. MIT Press.

Huang, C.-R., Chang, R.-Y., Lee, S.-B. (2004). Sinica BOW (Bilingual Ontological Wordnet): Integration of Bilingual WordNet and SUMO. Presented at the 4th International Conference on Language Resources and Evaluation (LREC2004). Lisbon. Portugal.

Lakoff, G. and Johnson, M. (1980). *Metaphor We Live By*. University of Chicago Press, Chicago.

Lakoff, G. (1993). The Contemporary Theory of Metaphor. In Ortony, A. (ed.). *Metaphor and Thought*. (second edition). Cambridge, CUP, pp. 202--251.

Niles, I. and Pease, A. (2001). Towards a Standard Upper Ontology. In the *Proceedings of the international conference on Formal Ontology in Information Systems*. (FOIS-2001). Ogunquit, Maine. pp. 2--9.

Wu, Z. and Palmer, M. (1994). Verb Semantics and Lexical Selection." In the Proceedings of the *32nd Annual Meeting of the Association for Computational Linguistics*. pp. 133--138.