

Source Domain Determination: WordNet-SUMO and Collocation

Siaw-Fong Chung

National Taiwan University
No.1, Sec. 4, Roosevelt Road
Taipei 106, Taiwan ROC
f91142002@ntu.edu.tw

Kathleen Ahrens

National Taiwan University
No.1, Sec. 4, Roosevelt Road
Taipei 106, Taiwan ROC
kathleenahrens@yahoo.com

Abstract

Conceptual metaphors provide in linguistic form the information that is mapped between two knowledge domains. The examination of conceptual metaphors has always involved a comparison of target and source mappings. However, defining what should be included in any one domain is a difficult issue. This paper claims that patterns in conceptual metaphor mappings can be found by first identifying the source domains through large-scale corpora analysis. Two methods are attested in this paper for source domains determination. These methods are top-down and bottom-up approaches. Top-down approach involves the use of knowledge domain ontology of SUMO (Suggested Upper Merged Ontology) and bottom-up approach involves the examination of collocation (through usage). Through using computational tools, the workability and precision of using these approaches will be compared.

1 Acknowledgements

We would like to thank Professor Chu-Ren Huang, Professor Yung-O Biq, Professor I-wen Su, Professor Hintat Cheung and Professor Sue J. Ker for commenting on this paper. We would also like to thank Professor Chu-Ren Huang's Shen-gen Project at Academia Sinica for supporting the discussion herein.

2 Introduction

Most studies of metaphors can be seen from two approaches, namely top-down and bottom-up. An example of top-down approach is using

knowledge domain ontology (Chung, Ahrens and Huang, 2005) by identifying domain information using taxonomy. The bottom-up approach, on the other hand, builds the knowledge domain through language use, i.e., by generating a pattern through analyzing how metaphors are used. This second approach has the underlying theoretical assumption of the prototype theory suggested by cognitive linguists such as Rosch and Mervis (1975), Labov (1973) and Wittgenstein (1978). This approach uses a frequency-based definition of prototypes. For example, the more frequently acceptable concept is the more prototypical concept.

This paper attests both top-down and bottom-up approaches through using computational tools such as WordNet and SUMO (top-down) and compares the results from the collocation method (bottom-up) in defining source domains. This is because the origin of metaphors can be explained through the systematic examination of the domains and the underlying reasons for source-target domain pairings (Heywood and Semino, 2005; Chung, Ahrens and Huang, 2005 and Mason, 2004). Therefore, the first way to uncover the origin of metaphors is through recognizing and identifying the domain information that is mapped.

3 Methodology

The target domain of *jing1ji4* 'ECONOMY' was discussed in Ahrens, Chung and Huang (2003) and other studies in the same series. However, these previous papers only discuss some of the source domains (such as PERSON, COMPETITION, TRANSPORTATION) and other source domains are not discussed in detail. Furthermore they do not incorporate collocation into their studies. The purpose of re-using this target domain is to re-evaluate the precision of the WordNet/SUMO methodology (in Chung, Ahrens and Huang, 2005) as compared to the collocation method as well as to examine the

other possible source domains that were not discussed.

3.1 WordNet-SUMO Method

The WordNet/SUMO methodology involves using metaphors extracted from the corpora manually. For the target domain of *jing1ji4* ‘ECONOMY,’ all instances were taken from the Sinica Balanced Corpus of Modern Chinese (<http://www.sinica.edu.tw/SinicaCorpus/>). The use of WordNet and SUMO is facilitated by the Sinica Bow interface (Huang et al., 2004). Sinica Bow provides the Chinese-English-Chinese translation of the WordNet senses as well as their related SUMO nodes. This interface allows one to look up Chinese senses and the ontological nodes related to these Chinese senses. The steps of using WordNet/SUMO method will be described in detail below.

For the search in Sinica Corpus, a maximum result (due to licensing limitation at the time of the search) of 2000 instances was collected. All the instances were analyzed manually and the metaphorical expressions were extracted.

Three hundred and thirteen metaphorical expressions were collected and these expressions were looked up in Sinica Bow using the Chinese-English look-up search engine. For all the Chinese words that were searched, the program returned the possible senses for them. From the list of the senses, the most concrete one was chosen manually. This is to find out the most possible concrete sense for this term to appear. The concrete sense then helps determine the possible source domain for a particular metaphorical expression. The following Table 1 is the summary of the information from the WordNet/SUMO method (for selected examples).

Table 1 Using WordNet/SUMO Method for *jing1ji4* ‘ECONOMY’

Metaphorical Expressions	WordNet Explanations	SUMO nodes
<i>qi3fei1</i> (take off)	take off from the ground, as of an aircraft or balloon	Transportation
<i>qing1lue4</i> (invasion)	the act of invading; the act of an army that invades for conquest or plunder	ViolentContest
<i>jian4she4</i> (construction)	X	

The suggested source domain for ‘take off’ is either AIRCRAFT or BALLOON whereas for ‘invasion’ is WAR.

For expressions that were not found in Sinica Bow (such as *jian4she4* in Table 1), their possible source domains could not be decided. This is one limitation of this method and it was suggested that the collocation method may be able to compensate when look-up fails.

3.2 Collocation

In order to obtain collocations for the Chinese metaphorical expressions, this paper uses the Chinese Sketch Engine (Kilgarriff, Huang, Rychly et al., 2005), which is a query system that sort concordance instances according to grammatical relations such as subject-of-query, object-of-query and modifies-query based on the Chinese Gigaword corpus. Its design follows the English Sketch Engine which is based on the British National Corpus. The English system is developed by Kilgarriff and Tugwell (2001).

For all metaphorical expressions extracted from corpora, all the Chinese terms were keyed in to the Chinese Sketch Engine system. For example, when *qi3fei1* ‘takeoff’ was entered in the Chinese Sketch Engine (with minimum one occurrence), the following search result in Figure 1 was found.

Since *qi3fei1* is an action, what we would like to know from the Sketch Engine is what other nouns that also take this verb. This information is captured when the collocates of *qi3fei1* occur at the subject position, i.e., when used literally, what kind of subjects will take the verb *qi3fei1*. In Figure 1 below, the frequency for *qi3fei1* is 12,758. The second column after the Chinese collocates shows the frequency of a particular collocate at a particular grammatical relation to the searched word. In the third column is the saliency value for the pair of collocation.

According to Kilgarriff and Tugwell (2001), “[s]aliency is estimated as the product of Mutual Information *I* (Church and Hanks, 1989) and log frequency.” However, Kilgarriff and Tugwell modify the Mutual Information value *I* by taking into consideration the overall frequency of the grammatical relation as compared to the other relations. The purpose of doing so is to avoid cases where low frequency collocates such as those which occur once but its mutual information value is high because it is the only time it appears together with the keyword. Therefore, the saliency value in the Chinese is a reliable calculator instead of the frequency value.

Based on this reason, the following discussion will take the saliency value as the main reference for the choice of collocates.

Figure 1

起飛 chinese_giga_trd freq = 12758

subject	2208	7.7	modifier	1467	4.3
飛機	514	59.05	無法	117	31.8
班機	225	47.16	再度	73	30.93
跑道	70	39.32	十分	69	30.73
經濟	576	36.77	按時	17	30.64
夢想	27	32.33	剛	44	30.59
客機	51	30.98	再	121	28.11
滑行道	7	28.4	正在	45	26.17
專機	25	25.69	重新	44	25.03
航機	14	24.63	即將	40	24.19
小時	32	23.05	剛剛	12	23.17
航空母艦	15	22.34	不能	43	23.06
航班	14	21.6	如期	15	21.79
包機	14	21.31	不准	12	21.56
軍機	15	21.07	全速	5	21.06
戰機	25	20.96	尙未	23	18.43
直昇機	17	19.84	再次	15	18.11

This paper selects the top six collocates for each metaphorical expression and analyze the possible source domains based on these collocates.

From Figure 1, four out of six collocates for *qi3fei1* are related to AIRPLANE ('airplane,' 'flight,' 'lane (for flight),' and 'customer flight'). Based on the previous results of WordNet-SUMO and collocation, the source domain for *qi3fei1* is suggested to be AIRPLANE (rather than BALLON).

On the other hand, most of the top collocates for *qing1lue4* 'invasion' are 'Japanese army,' 'militarism,' 'military affair' and 'weapon.' These collocates are found related to WAR and the source domain of WAR are selected.

Based on these two methods, the precision of defining a source domain is compared.

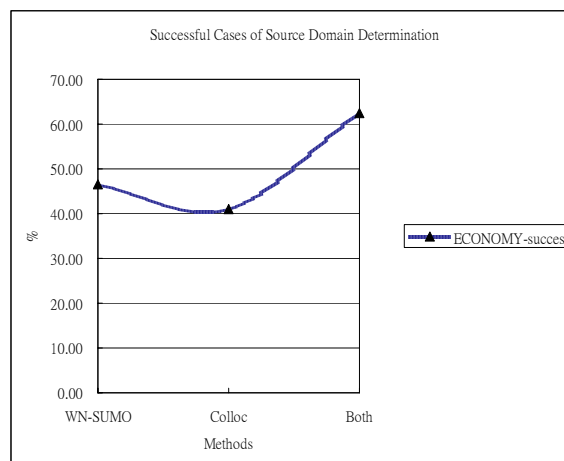
4 Comparing Precision

In order to compare the precision of using a) WordNet/SUMO only; b) collocation only; and c) both WordNet/SUMO and collocation, the percentages of the precision of successful or non-successful cases were calculated in below:

Number of successful/non-successful cases
----- x 100%
Total type of metaphorical expressions in a particular source domain

The results are given in Figure 2 below.

Figure 2



From Figure 2, one can see clearly that there is an obvious increase in percentages when both methods were used. The target domain of ECONOMY decreases slightly with the use of collocation. The reason why collocation did not work for ECONOMY is because the collocates for the metaphorical expressions are also metaphors. An example is the use of *cheng2zhang3* 'growth' where most of the collocates are metaphor such as 'economy,' 'career,' etc. Therefore, the possible source domains cannot be identified.

5 Conclusion

This paper suggests a way to compare the workability of top-down or bottom-up approaches in determining source domains. From the current analysis of 2,000 instances of *jing1ji4*, collocation method seems to be slightly less precise than the WordNet/SUMO method. However, there are several limitations that may affect this result. One of them is that the size of the data is not representative enough and second is that the manual determination discussed in the previous section still needs to be refined.

The proposal of using two linguistic approaches to analyzing conceptual metaphor has not been carried out before, as conceptual

metaphors are usually treated at the conceptual level. This paper, thus, provides empirical data for the issue. The results discussed herein will have theoretical implications as well as methodological contributions in terms of defining knowledge domains.

6 Future Work

The questions remaining for the paper involve how to operationalize the determination of the source domains. For example, in the WordNet/SUMO method, the steps based on intuition occur when a) selecting the most concrete WordNet senses from Sinica Bow; and when (b) determining the keywords in the WordNet definitions and SUMO nodes. For the collocation method, the selection of collocates that constitute a source domains from the six top collocates is still carried out manually. The aim for further research is to reduce the subjectivity of these steps and the following gives the possible ways of dealing with this issue.

In order to reduce the manual selection of the concrete sense from WordNet, one possibility is to take the SUMO node for each sense and compare their level of abstractness or concreteness in the SUMO hierarchy. Senses that fall under the abstract node are often not used literally. This is because the literal sense is usually more concrete. As for the collocation method, one of the possible ways to reduce the selection of concrete words is by searching for each of the collocates again in Sinica Bow to find out their related SUMO nodes. By doing so, one is able to find information of how these collocations are represented in the ontology. One can also use the upper hierarchies from WordNet to look at how one collocate relate semantically to another. For instance, *lu4xian4* 'route' and *lu4* 'road' may be synonyms and their relatedness can be seen from the WordNet hierarchies. By first establishing the relatedness of the collocates, one can avoid manually determining which of the collocates should be selected among the top six collocates.

Overall, the use of combination of method was not seen in previous work and by doing so; this work contributes not only to categorizing lexical items but also provides a platform for comparing the methodologies used in analyzing conceptual metaphors in corpora.

References

- Ahrens, Kathleen, Siaw-Fong Chung and Chu-Ren Huang. 2003. "Conceptual Metaphors: Ontology-based Representation and Corpora Driven Mapping Principles." In the *Proceedings of the ACL Workshop on the Lexicon and Figurative Language*. Sapporo, Japan. pp. 35-41.
- Chung, Siaw-Fong, Kathleen Ahrens and Chu-Ren Huang. 2005. "Source Domains as Concept Domains in Metaphorical Expressions." *Computational Linguistics and Chinese Language Processing (CLCLP)*. 10. pp. 553-570.
- Heywood, John and Elena Semino. 2005. "Source 'scenes' and source 'domains': Insights from a Corpus-based Study of Metaphor for Communication." Paper presented at the Third Interdisciplinary Workshop on Corpus-Based Approaches to Figurative Language, University of Birmingham, U.K., July 14, 2005.
- Kilgarriff, Adam, Chu-Ren Huang, Pavel Rychly, Simon Smith, David Tugwell. 2005. Chinese word sketches. In the *Proceedings of Asialex*, Singapore.
- Kilgarriff Adam and David. Tugwell. 2001. "WORD SKETCH: Extraction and Display of Significant Collocations for Lexicography." In the *Proceedings of the ACL Workshop COLLOCATION: Computational Extraction, Analysis and Exploitation*. Toulouse. pp. 32-38.
- Labov, W. 1973. "The Boundaries of Words and their Meanings." In Bailey C.-J. N. and R. W. Shuy. *New Ways of Analysing Variation in English*. Washington: Georgetown University Press. pp.340-373.
- Mason, Zachary J. 2004. "CorMet: A Computational, Corpus-Based Conventional Metaphor Extraction System." *Computational Linguistics*. 30. pp. 23-44.
- Rosch, E. and C. B. Mervis. 1975. "Family Resemblance: Studies in the Internal Structure of Categories." *Cognitive Psychology*. 7. 573-605.
- Semino, Elena. 2002. "A Sturdy Baby or a Derailing Train? Metaphorical Representations of the Euro in British and Italian Newspapers." *Text*. 22(1). pp. 107-139.
- Wittgenstein, L. 1978. *Philosophical Investigations*. Translated by G.E.M. Anscombe. Oxford: Basil Blackwell.