Ontology-based Prediction of Compound Relations — A study based on SUMO

Jia-Fei Hong
Institute of Linguistics
Academia Sinica
jiafei@gate.sinica.edu.tw

Xiang-Bing Li
Institute of Information
Science, Academia Sinica
dreamer@hp.iis.sinica.edu.tw

Chu-Ren Huang
Institute of Linguistics
Academia Sinica
churen@gate.sinica.edu.tw

Abstract

This paper explores the interaction between conceptual structure and morpho-syntax. In particular, we show that ontology-based conceptual classification can be used to predict internal relations in compounds. We propose an ontology-based approach to predict the semantic relation between the two component words in Mandarin VV compounds. A Mandarin VV compound is classified according to the eventive relation between the two simplex verbs. These relations specify how the eventive meanings of the two simplex verbs combine to form the meaning of the compound. The three types of eventive relations that we deal with in this paper are: coordinate, modificational, and resultative. Since the way in which two events combine with each other depends upon their event types, we hypothesize that the eventive relations can be predicted by the conceptual classified event types of the two simplex verbs. An approach of ontology-based prediction is proposed based on this hypothesis. The assignment of ontology classification for each simplex verb is based on SUMO and Sinica BOW. The correlation between the ontology class of each verb position and each eventive type is trained and scored based on a manually tagged lexical database. We encode the ontology information of each VV compound in a 3-tuple based on these correlation scores. This 3-tuple is represented as a three-dimensional vector and used to predict the eventive type of new VV compounds. Our classification experiment on unknown VV compounds yields good recall and precision.

1. Introduction

This paper explores the interaction between conceptual structure and morpho-syntax. In particular, we demonstrate that ontology-based conceptual classification can predict the internal relations in compounds. Even though the interaction between lexical semantics and syntax has been a

fertile ground for research in both theoretical and computational linguistics, there has been limited work on the interaction on semantics and morphology. This is somewhat unexpected since both lexical semantics and morphology are lexical operations. In Chinese, for instance, study on the formation of compound verbs has so far focused on the grammatical categories and grammatical functions of the simplex words being combined. However, it is clear that the lexical meaning of the simplex words must be preserved in compound formation. In other words, simplex words are building blocks of compounds that necessarily contain their grammatical information, including lexical meaning. Given the existence of the lexical meanings in the component words, an interesting theoretical issue is how they combine to form the meaning of the compound and to examine if the combination is governed by structural or conceptual principles?

In this paper, we deal with the internal relations of the VV compounds. Since the two components are the same lexical category, the internal relation is more likely to be conceptual. Our goal is to show that the different types of internal relations can be predicted with conceptual information. In particular, we want to show that ontological information can be effectively used. The ontology that we use is the Suggested Upper Merged Ontology (SUMO) developed by the IEEE upper ontology working group. We take different SUMO concept node combination for each compound verb and use it to predict the compound verb structure combinations.

2. Survey and motivation

Compound verbs attracted the attention of computational linguists because they demonstrated how grammatical form could interact with lexical semantics. The prediction of the syntactic and semantic relation between the two conjunctional verbs is one of the most challenging research topics. The challenge is even more intriguing in Chinese because compound verbs lack morphological markings (Chen and Hong 1995, Chen and Chen 1997 Chang and Chen 1999, Chang et al. 2000). This means that there is no overt information at all when a compound is formed with two words with identical categories, such as VV compounds.

There are five types of VV compound verbs in Modern Chinese according to Chao (1968): Coordinate, modifier-head, resultative verb, subject-predicate, and predicate-object construction. Of the five types, the first three types are composed of two verbs, hence referred to in the literature as VV compounds. Since the first three types have the same morphological composition, they must be differentiated by other means.

The identification of verbal compounds is a challenging yet essential task, as shown in Zackorva et al. (1999) for Czech. However, determination of the inter-relation of the two verbal elements poses an even harder challenge. McDonald (1995) showed that such predictions are not structure based and argues for a functional theory of the acquisition of compound order in English. Huang and Lin (1992) proposed an account for the prediction of the argument structure of VV compounds based on event templates. Drawing lessons from the above studies, we adopt the premise that the ordering of the two

verbal roots in a compound verb is determined by their eventive relation. We further assume that this eventive relation can be inferred based on the conceptual location of each verb. We propose to identify the conceptual location of the component verbs based on SUMO (Suggested Upper Merged Ontology). Each verb can be assigned to an ontology node through Sinica BOW (Bilingual Ontological WordNet). Generalizations about the eventive relations typical of each type of verbal compounds (coordinate, resultative, modificational) can then be made from corpus observations and the resulting generalizations are applied to predict the compound types.

3. The Data

3.1. Compound Verbs

There are 22,626 different (i.e. word types) compound verbs in the Sinica Corpus. For this study, the following forms are not included: First we exclude reduplicated compounds since the they are of identical form and the reduplication itself marks the meaning of being tentative, such as shang-shang 上上 'to go', xi-xi 習習 'to become used to', tao-tao 套套, 'to cover'. Second, we exclude stative verb constructions since these are adjectival in meaning, such as huan-le 歡樂 'to be happy', shu-shi 舒適 'to be comfortable'. Lastly, we exclude VN compounds; such as you-bing 有病, 'to be sick', shou-xian 收現, 'to take cash'. The remaining 19,496 types are classified into three types: coordinate, modificational, and resultive verb construction. There are 7,898 different coordinate compounds (40.51%), e.g. zhui-sha 追殺 'to pursue and kill', gong-shou 攻守 'to attack and defend', shou-qu 收取 'to collect', sou-bu 搜捕 'to track down and arrest', sou-cang 蒐藏, to 'search and collect', and chan-rao 纏繞 'to wind around'. Secondly, there are 6,498 (33.33%) different modificational verbs compounds, e.g. wei-xiao 微笑 'to smile', gai-chen 改稱 'to rename', nai-cao 耐操 'to withstand hardship', gai-kan 改看'to re-look', nan-ao 難熬 'to be not sufferable' and hao-zhan 好戰 'to be bellicose'. Finally, there are 5,100 (26.16%) different resultative verbs, e.g. da-si 打死 'to hit and kill', chi-bao 吃飽 'to eat to full', qiao-po 敲破 'to knock and break', chong-hun 沖昏 'to be over-conceited', la-duan 拉斷 'to pull and break', and bao-jin 抱緊 'to hold tight'.

3.2. Concept nodes mapping

In order to find out the conceptual composition of the VV compounds, we first identify the two component simplex verbs V1 and V2 for each VV compound. They are assigned conceptual nodes from SUMO. In addition, the internal relation of the VV compound has already been assigned to one of the three types: coordinate, modificational, and resultative. Hence, we now have the data of how a ontology class correlates with the internal relation classification for both the V1 and V2 position.

3.3. Data Analyses

We first calculated the distribution of the grammatical categories of the verbs in VV compounds as background information. The result is given in Table 1 below.

V	term frequency	%	V	term frequency	%	V	term frequency	%
VA	3663	15.78	VD	417	1.80	VHC	266	1.15
VAC	71	0.31	VE	923	3.98	VI	125	0.54
VB	541	2.33	VF	225	0.97	VJ	1158	4.99
VC	8829	38.03	VG	544	2.34	VK	294	1.27
VCL	741	3.19	VH	5314	22.89	VL	107	0.46

Table 1: POS distribution of verb types in VV compound

3.4. Compiling the training data

We manually analyzed 10% of all the data as the training set. In addition to the distributional data, we also found that:

(1) For the coordinate VV compound, the ten most frequent concept classes to occur at the V1 position are: *motion, process, getting, contest, giving, communication, attaching, touching, putting, and cutting.* The ten most frequent concept classes to occur at the V2 position are: *motion, process, attaching, giving, subjective assessment attribute, eating, putting, touching, removing, and intentional psychological process.*

(2) For the modificational VV compound, the ten most frequent concept classes to occur at the V1 position are: *subjective assessment attribute, intentional process, attaching, direction change, removing, positional attribute, group, connected, repairing, and unilateral getting*; The ten most frequent concept classes to occur at the V2 position are: *communication, motion, walking, seeing, expressing, process, eating, putting, music, and impacting.*

(3) For the resultative VV compound, the ten most frequent concept classes to occur at the V1 position are: *motion, removing, putting, cutting, process, impelling, eating, impacting, getting, touching*; The ten most frequent concept classes to occur at the V2 position are: *subjective assessment attribute, motion, origin, process, state of mind, destruction, attaching, near, organism process, and shape change.*

4. Experiments: ontology-based prediction of compound relations

4.1 Obtaining the Likelihood Scores

In order to empirically determine the correlation between simplex verb concepts and their compound relations, we randomly selected 2,400 VV compounds from our lexicon as the training data. That is, there are 800 compounds each belonging to the three types of relations. Each simplex verb (i.e. V1 and V2 for all 2,400 compounds) is assigned a suitable SUMO concept with manual verification. We observe the concept combination of these verbs in three structures, and reorganized the combination rules. The rules will be used to judge a verb which one structure it belongs.

Next, we calculated the likelihood score of a semantic relation S given the concept C from a simplex verb. The scores for each semantic relation: coordinative, adjunct-head, and resultative are calculated independently for both V1 and V2 positions. The likelihood score is obtained by the following formula.

$$L - Score_{c,s} = \frac{tf_{c,s}}{\sqrt{\sum_{1}^{n} (tf_{i,s})^2}}$$

L-Score$_{c,s}$: The likelihood score of C occurring in S for either V1 or V2

$tf_{c,s}$: The term frequency of C in S

$tf_{i,s}$: The term frequency of the *i*th concept in S

4.2 Experimental procedure

The testing set is collected by applying the unknown word detection tool of Ma and Chen (2003) on internet news articles. The randomly chosen compound verbs are manually tagged for the correct classification of their semantic relations. Each semantic relation type is given roughly equivalent number of testing examples: coordinate (92), modificational (91) and resultative (95).

For each verb, we assign its V1 and V2 to the suitable concept based on the verified testing data. However, if any particular verb does not have a verified a concept from the verified set, we use Sinica BOW (http://bow.sinica.edu.tw) or a Chinese electronic dictionary (Chen, etc., 1996) to determine its ontological concept assignment. We take three steps to determine that semantic relations classification of an unknown VV compound.

Step1: Identify all possible SUMO concepts for V1 and V2 for each unknown VV compound W, using the auxiliary language material if necessary.

Step2: Vectorize the correlation between the conceptual components of W and the three types of compound relations. The correlation vector of W is the sum of the correlation vector of V1 and the correlation vector of V2.

Step 2.1: The correlation vector of each component verb Vn is calculated as the sum of the likelihood score of all the possible ontology classifications for that particular verb, with each compound relation type assigned as one dimension. That is, for each V1 and for all its possible ontological node assignment, Wv1 = Wvci x.....x Wcn

Step 2.2: For each Vector Wvcn, the three dimensions are determined by their likelihood score for a semantic relation. For Wvcn <*x, y, z*>, x is the likelihood score for being a coordinate compound, y is the likelihood score for being a modificational compound, and z is the likelihood score for being a resultative compound.

Step3: Use the following formula to calculate the Cos(W, S), where Sc stands for coordinate, Sm stands for modificational, Sr stands for resultative. The highest Cos (W, S) value determines the relation type for compound VV.,

$$vectorW =< a_1, a_2, a_3 >, vectorS =< b_1, b_2, b_3 >$$

$$Cos(W,S) = \frac{\sum_{i=1}^{n} a_i \bullet b_i}{\sqrt{\sum_{i=1}^{n} a_i^2} \times \sqrt{\sum_{i=1}^{n} b_i^2}}$$

where $vectorSc =< 1,0,0 >, vectorSm =< 0,1,0 >, vectorSr =< 0,0,1 >$

Sc：The vector of coordinate compound

Sm：The vector of modificational compound

Sr：The vector of resultative compound

4.3 Experiment: Resolution and Prediction

4.3.1 Experiment with Basic Prediction Method

When assigning the likelihood scores, we need to resolve the multiple concept assignment to the verbs. Such assignment could either be attributed to the fact that the verb is ambiguous or that the verbal meaning may contain two or more meanings at the same time. In the first stage of our study, we do not attempt to resolve the ambiguity. Instead, the highest score for each verb is chosen, assuming that the more prominent relation is indeed the correct relation. The resultant recall and precision rates are listed in Table 2. Please note that the benchmark of random prediction among the three types of relations is 33.33%.

	Coordinate	**Modificational**	**Resultative**	**Average**	**All data**
Recall	32.65%	67.19%	66.10%	55.31%	51.58%
Precision	34.78%	47.25%	41.05%	41.03%	41.01%

Table 2: Recall and precision of basic experiment (single highest score)

The above result, although clearly superior to the default benchmark, has a lower than 50% precision rate and is not enough as an explanatory account. When we examine the assumptions made in this experiemnt, the assumption that the highest score is the correct one seems to be the most dubious. Hence, we revise that assumption and try to account for all possible correlations. In particular, we adopt a fractional voting system. That is, we tabulate all possible correlations and add up all their scores, and select the highest cumulative score (instead of the single highest score in the first study.) Table 3 shows the result of this revised method. Compares with Table 2, the vote mechanism clearly improves the precision of the prediction for both modificational and resultative compounds, while the improvement was less marked for the coordinate compounds. In the later experiments, we will only use the vote mechanism to predict the internal relation of the compound verbs.

	Coordinate	**Modificational**	**Resultative**	**Average**	**All data**
Recall	35.42%	62.65%	64.00%	54.02%	52.76%
Precision	36.96%	57.14%	50.53%	48.21%	48.20%

Table 3: Recall and precision of basic experiment with the fractional voting

4.3.2 Incorporating Heuristic Knowledge

One thing in common in both Table 2 and Table 3 is the lower than average recall and precision of the coordinate compounds. This can be explained since the coordinate compound, unlike the other two types, depends less on the conceptual classification of either V1 or V2 but rather than the collocation of V1 and V2. Hence we look for heuristics that would help to differentiate the coordinate compounds from the others. We first observe that there is a substantial number of modificational compounds that are wrongly predicted to be coordinate compounds. For most of these wrongly predicted cases, the POS of V1's is VH (i.e. stative verb, roughly equivalent to English predicative adjectives). In other words, we can generalize that compounds with a VH as V1 must be modificational, not coordinate. The improved result with this added heuristics is given below:

	Coordinate	**Modificational**	**Resultative**	**Average**	**All data**
Recall	39.08%	61.86%	63.38%	54.77%	54.51%
Precision	36.96%	65.93%	47.37%	50.09%	50.00%

Table4. The improved result after dealing with the pos of V1 (VH)

Table 4 shows a marked increase of over 8% in the prediction of modificational compounds, as expected. What is interesting is that the recall and precision of coordinate compounds also improve.

Another type of heuristic knowledge that will be useful is the lexical knowledge of specific verbs. In particular, the resultative compounds are predominant among a very particular subset of verbs in Mandarin. In particular, Chiu, Luo and Chen's (2004) study identify a set of V2's that can be reliably

predicted to be part of a resultative compound. There are two possible ways to use this lexical knowledge. On one hand, we can take this as positive evidence. That is, VV compounds whose V2 belongs to the set are classified as resultative compounds. We can also take this as negative evidence. That is, those V2s that do not belong to the list are not resultative compounds. The result with the positive evidence assumption is given first in Table 5.

	Coordinate	Modificational	Resultative	Average	All data
Recall	66.67%	67.90%	70.54%	68.37%	68.99%
Precision	34.78%	60.44%	95.79%	63.67%	64.03%

Table5: The result for adding to deal with the list of resultative V2 positively

As expected, Table 5 shows an increase in both the recall (+26%) and precision (+48%) of resultative compounds, while precision of VV and modifier V is reduced. Recall of coordinate compounds also improved. But there is also some decrease in the precision of the prediction of the coordinate and modificational compounds.

In the next and last experiment, the lexical knowledge of resultative V2's is used as negative evidence. That is, if the verb V2 is not among the attested list of resultative verbs, we assume the compound verb is not likely to be a resultative verb. The result is given in Table 6.

	Coordinate	Modificational	Resultative	Average	All data
Recall	67.16%	67.06%	83.02%	72.41%	73.64%
Precision	48.91%	62.64%	92.63%	68.06%	68.35%

Table 6: Recall and precision of the revised experiment with list of resulative verbs used negatively

There is 14% improvement in the precision of VV structure. The recall rate is now over 67% for all three types and the precision ranges from just under 50% for coordinate compounds to over 92% for resultative compounds. This last experiment showed marked improvement for both coordinate compounds and resultative compounds by addressing the previous mis-classification between coordinate and resultative compounds.

5. Conclusion

Our study shows that the conceptual information expressed in terms of ontology classification can be used to predict the semantic relations within VV compounds. In particular, we show that the concept-based prediction can integrate other knowledge sources and yield good prediction results. This point should be highlighted since the semantic information that a human being uses in language processing is not limited to an abstract ontology. Hence, it is probably more important to show that an

ontology based system can be greatly improved when other information sources are integrated, rather then showing that ontology alone can have very high rates of prediction. In other words, ontology is the infrastructure of knowledge which does not claim to have enough knowledge by itself, but instead provides a base for incremental integration of other knowledge sources.

It is important to note that our current study made the assumption that the conceptual classes of the two simplex verbs V1 and V2 can be used to predict the compound internal relation. This assumption is largely attested by our study. However, we also noted that our study yielded the least satisfactory results for the coordinate compounds. This is because coordinate compounds rely crucially on the compatibility of V1 and V2, which is not modeled in the current work. We will incorporate additional mechanisms to deal with conceptual compatibility of the two elements in the future.

Acknowledgements

We would like to thank all members of the Chinese Knowledge Information Processing Group, especially Professor Keh-jiann Chen and colleagues of the CWN group, for their help and comments. Thanks also go to Professor Kathleen Ahrens for commenting on an earlier version of this paper. All remaining errors, of course, are ours.

References

Chao, Y.R.1968. A grammar of spoken Chinese. University of California Berkeley, California.

Chang, Li-li, Chen, Keh-Jiann and Huang, Chu-Ren. 2000. A Lexical-Semantic Analysis of Mandarin Chinese Verbs: Representation and Methodology. Computational Linguistics and Chinese Language Processing. Vol.5, No. 1, February 2000, pp. 1-18.

Chang, Li-li, Chen, Keh-Jiann. 1999. 動詞詞構與語法功能互動初探. Proceedings of the 12th Research on Computational Linguistics, Tsinchu, Taiwan. pp. 67-85

Chen, Keh-Jiann, Chen, Chao-Jan. 1997. A Corpus-based Study on Computational Morphology for Mandarin Chinese. Round Table Conference: Quantitative and Computational Studies on the Chinese Language, Hong Kong. pp. 283-306.

Chen, Keh-Jiann, Huang, Chu-Ren., Chang, L. P., & Hsu, H. L., 1996, "Sinica Corpus: Design Methodology for Balanced Corpora," in Proceedings of PACLIC II, Seoul, Korea, pp. 167-176.

Chen, Keh-Jiann, Hong, Wei- Mei. 1995. 中文裡「動ㄧ名」述賓結構與「動ㄧ名」偏正結構的分析. Proceedings of the 8th Research on Computational Linguistics. pp. 1-13.

Chiu C.M., Luo Ji-Qing, Chen Keh-Jiann. 2004. 現代漢語複合動詞之詞首詞尾研究, ROCLING XVI: Conference on Computational Linguistics and Speech Processing, Taipei, Taiwan, ROC. pp131-139.

Eva Zackova, Lubos Popelinsky and Milos Nepil. 1999. Recognition and Tagging of Compound Verb Groups in Czech. NLP Laboratory, Faculty of Informatics, Masaryk University.

Huang, Chu-Ren, Lin, Fu-Wen. 1992. Composite Event Structures and Complex Predicates: A Template-based Approach to Argument Selection. Proceedings of the Third Annual Meeting of the Formal Linguistic Society of Midamerica (FLSM III). Bloomington: IULC. pp. 90-108. Z. Wu and M. Palmer "Verb Semantics and Lexical Selection".

Ma Wei-Yun & Keh-Jiann. 2003. A bottom-up Merging Algorithm for Chinese Unknown Word Extraction. Proceedings of ACL workshop on Chinese Language Processing, pages 31-38

McDonald Scott. 1995. Learning Compound Order: Towards a Functional Explanation Center for Cognitive Science. University of Edinburgh. Edinburgh, Scotland.

Niles, I. and Pease, A. 2003. Linking Lexicons and Ontologies: Mapping WordNet to the Suggested Upper Merged Ontology. In Proceedings of the 2003 International Conference on Information and Knowledge Engineering, Las Vegas, Nevada.