View metadata, citation and similar papers at core.ac.uk
brought to you by CORE
provided by Waseda University Repository

Bridging the Gap between Graph Modeling and Developmental Psycholinguistics

An Experiment on Measuring Lexical Proximity in Chinese Semantic Space

Shu-Kai Hsieh[a], Chun-Han Chang[a], Ivy Kuo[a], Hintat Cheung[b], Chu-Ren Huang[c], and Bruno Gaume[d]

[a] National Taiwan Normal University / Academia Sinica
[b] National Taiwan University
[c] The Hong Kong Polytechnic University / Academia Sinica
[d] Institut de Recherche en Informatique de Toulouse

Abstract. Modeling of semantic space is a new and challenging research topic both in cognitive science and linguistics. Existing approaches can be classified into two different types according to how the calculation are done: either a word-by-word co-occurrence matrix or a word-by-context matrix (Riordan 2007). In this paper, we argue that the existing popular distributional semantic model (vector space model), does not adequately explain the age-of-acquisition data in Chinese. An alternatively measure of semantic proximity called PROX (Gaume et al, 2006) is applied instead. The application or PROX has interesting psycholinguistic implications. Unlike previous semantic space models, PROX can be trained with children's data as well as adult data. This allows us to test the hypothesis that children's semantic space approximates the target of acquisition: adult's semantic space. It also allows us to compare our Chinese experiment results with French results to see to attest the universality of the approximation model.

1 Introduction

Semantic space modeling has been an interesting topic both in cognitive science and linguistics. There have been many empirical methods proposed both in computational linguistics and in (cognitive) psychology. Recently, distributional models (or vector-based models) becomes one of the core techniques for their impact on advanced tasks such as modeling of human language processing and acquisition. Notwithstanding its significant success, we will argue in this paper, that this model exposes some crucial limitations when processing with Chinese data, especially in the context of psycholinguistic settings. Alternatively, we propose to take graph-based model to perform the task, a stochastic method (PROX) proposed by Gaume is tested on age-of-acquisition data in Chinese. The preliminary experiment yields promising results, and we believe that it will shed light on the cognitive modeling and bridge the gap between graph method and psycholinguistics.

This paper is organized as follows: Section 2 will critically review related works on semantic space modeling, including the introduction of the existing models and their limitations; then Section 3 discusses our age-of-acquisition data, how they are elicited to be linguistic evidences of early lexicon acquisition. Section 4 introduces the proposal mathematical method in measuring the lexical proximity and semantic approximation based on our data, and some incidental issues. Comparison and discussion will be shown in Section 5, finally Section 6 concludes this work.

2 Related Works on Semantic Space models

This section reviews the most popular approaches, - the distributional model -, in modeling semantic space: the methods, evaluation and problems.

Copyright 2009 by Shu-Kai Hsieh, Chun-Han Chang, Ivy Kuo, Hintat Cheung, Chu-Ren Huang, and Bruno Gaume

2.1 Distributional models

The hallmark of current distributional semantic models is that they all assume the *Distributional Hypothesis*, which states the surrounding context of a given target word provides information about its semantic content (Sahlgren, 2006). (Riordan and Jones, 2007) classifies semantic space models into two types based on their design architecture and representation scheme, i.e., "paradigmatic" and "syntagmatic" spaces. The representative paradigmatic Space modeling are *Hyperspace Analogue of Language (HAL) framework* (Burgess and Lund (2000)), which yields better results on the synonymy test. The representative syntagmatic space modeling, on the contrast, is *latent semantic analysis* (LSA; Landauer and Dumais, 1997), and other lexical co-occurrence models of semantic memory, which seems to have a higher degree of correlation with the human association norm.

Although the models differ considerably in the algorithms used, they are all fundamentally based on the principle that a word's meaning can be induced by observing its statistical usage across a large sample of language. Take the most popular model Latent Semantic Analysis for example, it is a statistical computational model which can automatically generate semantic similarity measures between words in a corpus of texts. It begins with constructing *vector spaces*, that is, by finding the frequency of terms used and the number of co-occurrences in each documents throughout the corpus, and in order to reduce the dimensionality of the constructed co-occurrence matrix, a projection method (e.g., Singular Value Decomposition (SVD)) is used to find deeper meanings and relations among words. The distributional model benefits from training on extremely large amounts of data, and aim to correlate more closely with human semantic similarity.[1]. The evaluation typically proceeds by first training a model on a corpus of text, after which the model can generate similarity ratings between word pairs for comparison with human judgments. [2]

2.2 Problems

In addition to some other known limitations of distributional models, such as treating *contexts* equally, ignoring the facts that using different data set and tuning parameters will affect the resulting semantic space, ..etc, there should be more methodological discussions and cross-cultural empirical tests on proposing these models as plausible simulations of human semantic space organization. In the following, we summarize two main issues we need to deal with:

First, to the best of our knowledge, current construction and evaluations of distributional lexical semantic models (e.g., LSA, Topics, HAL) have largely focused on word-level. However, analyzing performance at the word-level (such as synonym and word association tests) may provide some interesting insights into the behavior of the models, it does not capture important linguistic phenomena in Chinese, as wordhood is a continually debating issue. Technically, corpus preprocessing strategy affects the target and context selection. The evident effect on LSA modeling can be tested from different pre-processed corpus data (i.e., with different segmented texts), the space model thus built is prone to lose its cognitive plausibility.[3]

Second, as known that human lexical semantic competence is dynamically multi-faceted, but the evaluation of distributional models tend to focus on a single aspect (most typically the detection of semantic similarity), and mostly trained on adult input.

This misunderstanding may go back to the *metaphorical usage* of mental lexicon (Jarema and Libben). Mental lexicon as a metaphor implies a dictionary-like thing represented in our mind,

[1] Meanwhile, there have been other variants proposed, e.g. COALS (D.Rohde et al, 2009), that is inspired by the HAL and LSA methodologies

[2] Proposed (shared) tasks from ESSLLI workshop 2008 include: (1)Semantic categorization - distinguishing natural kinds of concrete nouns, distinguishing between concrete and abstract nouns, verb categorization, (2) Free association - predicting human word association behavior;(3) Salient property generation - predicting the most salient properties of concepts produced by humans.

[3] Interested readers can refer to the Chinese LSA website `www.lsa.url.tw/modules/lsa`.

which with its huge and growing storage of usages, allow us to engage in everyday processes of language comprehension and production. Yet, as the vast majority of psycholinguistic research show, it is rather a cognitive system that constitutes the capacity for conscious and unconscious *lexical activity*. The term cognitive system highlights the fact that, while making no claims regarding the extent to which the mental lexicon is monolithic and the extent to which it is structurally or functionally encapsulated, the mental lexicon is not as that entity *enables* lexical activity but, rather, as that entity which *is* lexical activity. In addition, from the viewpoint of language acquisition, the lexical component changes the most over the lifespan, with the acquisition of new words extending throughout adulthood. In that sense, our mental lexicons are never fixed and never cease being linguistic capacities. Taking these into considerations, there is room for improvement for the existing distributional models.

3 Age-of-Acquisition Data from M3 Projects[4]

3.1 Experimental Design and Results

One of the goals in our on-going M3 project is to show the importance of *semantic flexibility* during early lexical development (Duvignau, K. et al. 2004). In order to elicit the production of semantic approximation, an action-video naming task followed by a reformulation task has been performed modeled on previous experiments from our French colleagues. 17 action movies[5] were first presented in random order to each participant, the instructions were given at the time the action in the movie was completed and its results were visible (e.g., when the glass is broken). At that moment a question was asked to the participant: what did the woman do?[6] In this way 954/5769 (type/token) Chinese verbs were elicited. Figure 1 depicts the scenario.

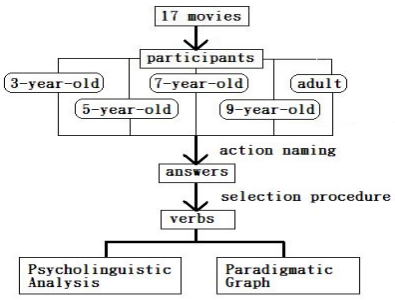

Figure 1: M3 project scenario

In order to perform the analysis of semantic flexibility, (Duvignau, 2002; Duvignau et al, 2005) distinguished two kinds of semantic approximations: (1) Intra-domain proximity or pragmatic approximation, which reflects synonymy between verbs that belong to the same semantic domain. E.g., a child with three years old uses the verb *coupe* ('cut') to designate the action of '*peels* an orange'.(2) *Extra-domain proximity* or linguistic approximation, which reflects synonymy between verbs that belong to distinct semantic domains. E.g., a child with 5 years uses *undress* to designate the action of '*peels* an orange'.

[4] M3 project is an international collaborative interdisciplinary research project that aims to survey the semantic verbal approximations in both French and Chinese mandarin early lexicon acquisition.

[5] break_glass, tear_off paper, peel_banana, saw_plank, cut_bread, peel_orange, ...etc.

[6] For more details please refer to M3 project website. `http://140.112.147.149:81/m3/`

4 Measuring Lexical Proximity and Semantic Approximation

4.1 Graph-based model

Given the data, our concerns turn to the interdisciplinary issue: what would be the best computational model that maximally approximate the semantic space represented in the early verb acquisition data?

As mentioned, corpus-based distributional semantic models treat semantic content as vectors[7], and in this way turns *distributional similarity* to *semantic similarity* through statistic analysis. However, in the context of semantic space modeling, the distinction of **similarity** and **proximity** should be carefully distinguished. As mentioned by (Gaume et al, 2006), that proximity is a Gestalt principle of network organization. In our attempt, measuring lexical proximity means calculating *social nearness* of lexicon, which should better be measured by considering the 'global' position in a semantic space, not solely by 'local' comparison of two given verbs.

4.2 PROX: Graph-based methods of Proximity Measurement

(Gaume et al., 2006) proposes 'proxemy', a semantic proximity measure based on the ground of the complete graph. PROX(PROXemy) is a stochastic method designed for studying "Hierarchical Small Worlds" (2008). The goal is set to measure and visualize proximity between lexical nodes. To sum, PROX build a similarity measure between the vertices, It takes a graph as input and transform them in a Markov chain whose states are graph vertices. The underlying hypothesis is that ares having a high density in edges correspond to closely related verb meanings (in a graph of verbs).

Formal Definition of PROX algorithm is given as follows:

Definition 1. Given a graph with n vertices, $G = (V, E)$, we will note $[G]$ the matrix $n \times n$ such that $\forall r, s \in V$, $[G]_{r,s} = 0$ if $\{r, s\} \notin E$ and 1 otherwise. $[G]$ is called the adjacency matrix of G.

Definition 2. Given $G = (V, E)$ a reflexive graph with n vertices. $[\hat{G}]$ is a $n \times n$ matrix defined by $\forall r, s \in V$, $[\hat{G}]_{r,s} = \frac{[G]_{r,s}}{\sum_{x \in V}\{[G]_{r,x}\}}$, $[\hat{G}]$ is the Markovian matrix of G.

Definition 3. Given $G = (V, E)$ a reflexive graph with n vertices and $[\hat{G}]$ its Markovian matrix, $\forall r, s \in V$, $\forall t \in \mathbb{N}^*$, PROX (G, t, r, s) = $[\hat{G}^t]_{r,s}$

PROX (G, t, r, s) is therefore the probability for a particle departing from r at the instant zero to be on s at the instant t. **When PROX (G, t, r, s)** $\succ$ **PROX (G,t,r,u)**, the particle has more probability to be, at instance t on s than on u and it is graph structure that determine these probabilities.

4.3 PROXing Chinese Graph Data

Table 1 shows the numeric data of Chinese graph data we processed. During the period of preparation, we have tried to explore different possibilities and make the most use of merging existing Chinese lexical resources, including Chinese Wordnet (CWN), CILIN Thesaurus (CILIN), Sinica Bilingual Ontological Wordnet (SinicaBOW), and English-Chinese Translation Equivalence Database (ECTED).

A crucial problem encountered while PROX Chinese Graph data, which is similar to the segmentation issue discussed previously, lies in that some 'verbal constructions' exist in our acquisition data do not constitute 'words'. Usually, resultative compound and resultative construction have semantic relation with the head word, our strategy here is to create a set of paradigmatic links to bypass the problem. (See Figure 4.3 A is C to B, A is E to D)

[7] Alternative names for DSMs:corpus-based semantics,statistical semantics,geometrical models of meaning,vector semantics word (semantic) space models, etc

Table 1: Numeric data of Chinese graph data

	Number of vertices	Number of edges	Number of arcs *
CILIN	30416	280002	529588
CILIN + CWN	30618	478809	927000
CWN	1970	5056	8142
CWN (without ID)	1885	4943	8001
ECTED	10019	20854	31689
ECTED + CWN	10706	40330	69954
SinicaBOW + CWN (Wordnet ID)	10661	40314	69967
SinicaBOW + CWN (Lemma Form)	9998	20876	31754

* including reflexive and symmetric links

A	B	C	D	E
分一半	分	TROPONYMY		
分成一大一小	分	TROPONOMY	分成	TROPONYMY
分開來	分	TROPONYMY	分開	SYNONYMY
切一切	切	SYNONYMY		
切切切割	切	SYNONYMY	割	SYNONYMY
打開來	打開	SYNONYMY		
打壞	打	TROPONYMY		
用一個洞	用	TROPONYMY		
用成一節一節的	用	TROPONYMY	用成	TROPONYMY
削破	削	TROPONYMY		
削起來	削	?		
K	打	SYNONYMY		
褪下	脫下	SYNONYMY		
嚕	推	SYNONYMY		

5 Comparison and Discussion

5.1 Comparing PROX with M3 Age-of-Acquisition data

For the film (`peel banana`), the elicited verb from adults is 剝 consistently, while a lists of different verbs from children aging from 3 to 9 are 切,扒,用,吃,弄,拉,拔,剝,拆,削,拿,剪,帶,脫,掰,撕. The distribution pie charts are plotted in Figure 2. For the purpose of comparison, we first run the LSA model[8] with the input 剝, calculated based on Academia Sinica Balanced Corpus. From the results (Figure 3), surprisingly none of verb retrieved is found in our list. We then run the PROX based on CILIN and ECTED-CWN data[9], the matched verbs (扒、用、拔、削、脫) yield the coverage of 30%. Figure 4 and 5 show the visualization of the semantic space around the target word and its neighbors.

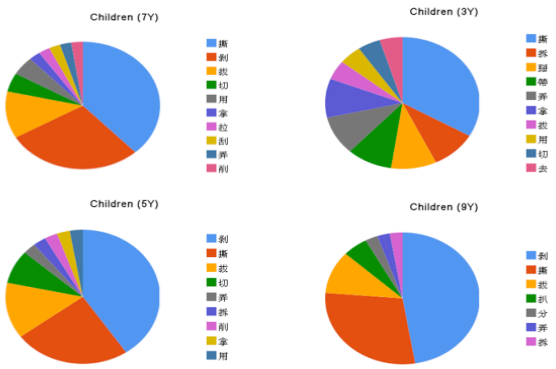

Figure 2: Age-of-acquisition verb data elicited from the film (”peel banana”)

5.2 Developmental issue

At a wider angle, we observed that, from the preliminary experiment, PROX model seems to perform better at semantic approximation tasks. However, it also raises the issue with regard to

[8] http://www.lsa.url.tw/modules/lsa
[9] http://erss.irit.fr:8080/graph-dev/

Results:

1 肚子 0.445 [du4 zi5] belly/abdomen/stomach
2 肉 0.433 [rou4] meat/flesh
3 雞 0.384 [ji1] fowl/chicken
4 倒 0.367 [dao3] to fall/to collapse/to topple/to change/to fail/bankrupt
5 放 0.362 [fang4] to release/to free/to let go/to put/to place/to let out
6 跑到 0.348 []
7 了 0.347 [le5] (modal particle intensifying preceding clause)/(completed action marker)
8 面前 0.343 [mian4 qian2] in front of/facing/presence
9 吃 0.336 [chi1] eat/eradicate/destroy/receive
10 剩下 0.329 [sheng4 xia4] to be left (over)/to remain
11 這時 0.327 []
12 大聲 0.322 [da4 sheng1] loudly/loud

Figure 3: Calculating 剝 with LSA model based on ASBC, with dimension 300, threshold 0.3

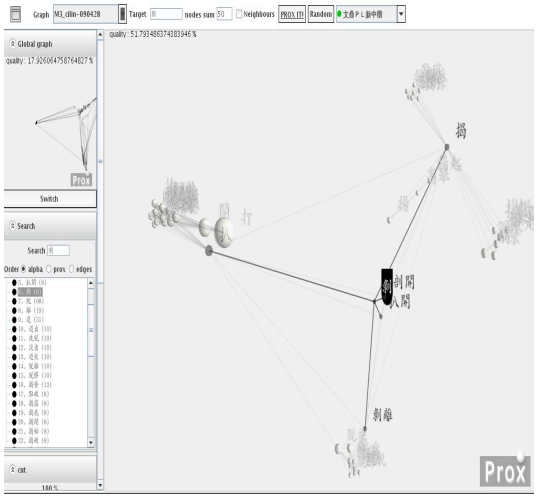

Figure 4: CiLin 剝, we set the value of t to 3 since L is less than 3 in the Graph. That is, we take into account the global graph simply by calculating PROX (G;3;r;s).

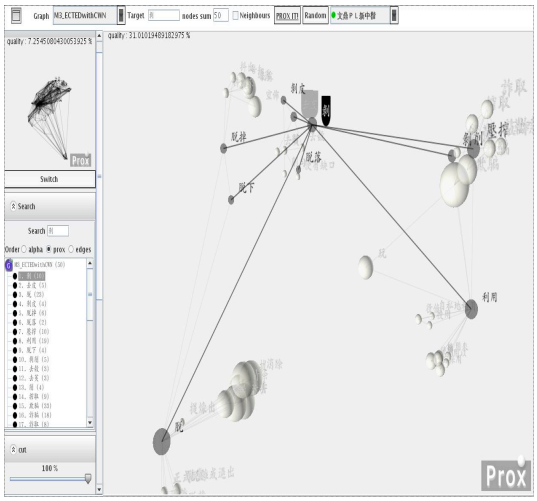

Figure 5: ECTED+CWN 剝

modeling the developmental changes in the mental lexicon. For instance, one of the distinguishing features of our experimental data lies in the age variable. Out of the seventeen movies, there are four movies where the children utterances show gradual adoption of the adult conventional usage. In each of them we extract the cross-age group highest frequency verb ($V1_{freq}$.), which is also the adult conventional utterance, to compare with the second highest frequency verb ($V2_{freq}$.). Both the $V1_{freq}$. and the $V2_{freq}$. occur in each age group of the respective movies, showing their close semantic relation. The original five age groups are regrouped into three to increase data amount of each group and hence reliability, with 3-year-old group and the 5-year-old put in one group and the 7-year-old grouped with the 9-year-old. Figure 6 demonstrates an example from the movie 'sawing plank'.

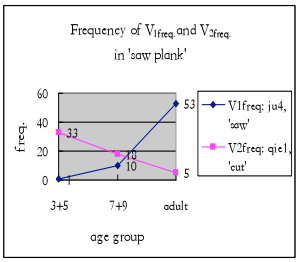

Figure 6: Frequency of $V1_{freq}$ and $V2_{freq}$ in 'saw plank'

Frequency of $V1_{freq}$ and that of $V2_{freq}$ by age group are shown in Figure 6. It is noticeable that while the frequency of $V1_{freq}$ increases over age, that of $V2_{freq}$ decreases on the contrary. We divide the frequency of $V2_{freq}$ by that of $V1_{freq}$ to produce what we call Replacing Rate, which is defined as Replacing Rate = Frequency of $V2_{freq}$ / Frequency of $V1_{freq}$. By replacing we mean that children who have not picked up the conventional verb replace it with an alternative verbal form. Figure 7 demonstrates the replacing rates in the four movies.

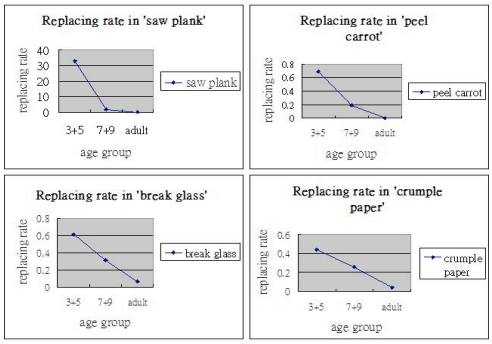

Figure 7: Replacing rates in the four movies

The replacing rates all show a clear drop over age. This means the replacing trend decreases along children's development when they gradually learn the adult conventional verb. These also hint something similar to what is called *pruning of neural connections* when describing children's neuro-cognitive development. While significant, this has been neglected in the study of computational modeling of the mental lexicon.

5.3 Toward a Cross-language and Multi-facet Testbed

The result of our initial experiment motivates us to work on preparing and evaluating different vector-based and graph-based models on different kinds of data, such as data from corpus (Academia Sinica Balanced Corpus, LDC Chinese Gigaword Corpus, Web as Corpus), lexical

resources (Chinese Wordnet, Cilin Thesaurus, wiki), psycholinguistic experiments (word association, priming, age-of-acquisition). With an integrated testbed, it would be interesting to evaluate how amenable various kinds of models are for which kinds of data, and validated by cross-lingual evidence.[10]

6 Conclusion

The aim of this paper is to explore the modeling and measurement of lexical proximity in Chinese semantic space. We argue that the existing popular distributional semantic model (also known as vector semantics model), does not adequately explain the age-of-acquisition data in Chinese. An alternatively measure of semantic proximity called PROX (Gaume 2004, 2006) is applied instead. Starting from our preliminary experimental results, we believe a large-scale cross-linguistic and multi-facets testbed is urgently needed. This constitutes one of our future works.

On the other hand, this paper represents a small step toward a larger understanding of the issues in modeling developmental changes of mental lexicon. This larger understanding of the issues require not only sophisticated insights from theoretical perspectives, which includes overgeneralization of semantic meaning in early lexical development: moon for ball (Clark 1993), and suggestions for lexical reorganization (Bowerman et al, 2004), but also novel techniques merging with mathematical models, computer simulations and empirical data analysis. The insights will illuminate the nature of human cognitive system.

References

Clark, Eve V. 1993. *The lexicon in acquisition*. Cambridge University Press.

Duvignau, K. 2002. *La métaphore berceau et enfant de la langue.* Ph.D. thesis, Université Toulouse Le mirail.

Duvignau, K. and B. Gaume. 2004. Linguistic, psycholinguistic and computational approaches to lexicon: For early verb-learning. Special issue on "learning". *Cognitive Systems*, 6-2(3).

Gaume, B., K. Duvignau and M. Vanhove. 2004. Semantic associations and confluences in paradigmatic networks. 2008. In: M. Vanhove ed., *From polysemy to semantic change: towards a typology of lexical semantic associations*. John Benjamins Publishing Company.

Gaume, B., F. Vernant and B. Victorri. 2006. Hierarchy in lexical organization of natural languages. *Hierachy in natural and social networks*, Methodos series, vol 3. Springer.

Johnson, M., Y. Munakata and R. O. Gilmore. 2002. *Brain development and Cognition : A reader*. Blackwell Publishers.

Majid, A., M. Bowerman, S. Kita, D. Haun and S. Levinson. 2004. Can language restructure cognition? The case for space. *Trends in Cognitive Sciences*, 8(3),108–114.

Landauer, T.K., P.W. Foltz and D. Laham. 1998. Introduction to latent semantic analysis. *Discourse Processes.* 25, 259-284.

Mehler, A. and L. Sichelschmidt. 2006. Re-conceptualizing latent semantic analysis in terms of complex network theory. A corpus-inguistic approach. *DGKL 2006*.

Riordan, B. and M. N. Jones. 2007. Comparing semantic space models using child-directed speech. In D.S.MacNamara and J.G.Trafton (eds), *Proceedings of the 29th Annual Cognitive Science Society*.

Sahlgren, M. 2006. *The word-space model*. Ph.D Dissertation, Stockholm University.

[10] Preparation website: http://140.112.147.149:81/m3/paclic_index.asp