IJCLA VOL. 1, NO. 1-2, JAN-DEC 2010, PP. 153-168
RECEIVED 24/11/09 ACCEPTED 16/01/10 FINAL 11/03/10

Identifying Different Meanings of a Chinese Morpheme through Latent Semantic Analysis and Minimum Spanning Tree Analysis

BRUNO GALMAR, JENN-YEU CHEN

National Cheng Kung University, Taiwan

ABSTRACT

A character corresponds roughly to a morpheme in Chinese, and it usually takes on multiple meanings. In this paper, we aimed at capturing the multiple meanings of a Chinese morpheme across polymorphemic words in a growing semantic micro-space. Using Latent Semantic Analysis (LSA), we created several nested LSA semantic micro-spaces of increasing size. The term-document matrix of the smallest semantic space was obtained through filtering a whole corpus with a list of 192 Chinese polymorphemic words sharing a common morpheme (港 gong1). For each of our created Chinese LSA space, we computed the whole cosine matrix of all the terms of the semantic space to measure semantic similarity between words. From the cosine matrix, we derived a dissimilarity matrix. This dissimilarity matrix was viewed as the adjacency matrix of a complete weighted undirected graph. We built from this graph a minimum spanning tree (MST). So, each of our LSA semantic space had its associated MST. It is shown that in our biggest MST, paths can be used to infer and capture the correct meaning of a morpheme embedded in a polymorphemic word. Clusters of the different meanings of a polysemous morpheme can be created from the minimum spanning tree. Finally, it is concluded that our approach could model partly human knowledge representation and acquisition of the different meanings of Chinese polysemous morphemes. Our work is thought to bring some insights to the Plato's problem and additional evidence towards the plausibility of words serving as ungrounded symbols. Future directions are sketched.

1 INTRODUCTION

Polymorphemic Chinese words are composed of the binding of two Chinese characters (e.g. 王公) or more (e.g. 公路赛) . We proposed a computational approach to extract the different meanings of 公 in a list[1] of 192 polymorphemic 公 words which occur in a corpus.

A Chinese character like 公 corresponds roughly to a morpheme in Chinese, and it usually takes on multiple meanings. For example, an etymology dictionary offers the following 16 senses[2] -16 etymological dimensions of meaning- for the character 公 (gong1) :

unselfish / unbiased / fair / to make public / open to all / public / the first of old China's five-grades of the nobility / an old Chinese official rank / the father of one's husband (one's husband's father) / one's father-in-law / one's grandfather / a respectful salutation / the male (of animals) / office / official duties / a Chinese family name

公 can take one of these meanings in the words in which it occurs. In the word 公平 (fair) the meaning of 公 is fair. In this case, the meaning of the morpheme is identical with the one's of the bimorphemic word. This *"fair"* meaning of 公 is different from the meaning of 公 in 公园 (public park,park) which is *"public, open"*.

Our computational approach to infer the meaning of 公 in polymorphemic words can be unfolded in five steps:

1. Through filtering a Chinese corpus by three nested list of words, we created three nested term-document matrices, weighted them and computed reduced Singular Value Decomposition (SVD) on them to obtain three nested Latent Semantic Analysis (LSA) semantic spaces.
2. For each LSA semantic space we computed the cosine matrix and the dissimilarity matrix for all terms.
3. We used each dissimilarity matrix as the adjacency matrix of a complete weighted undirected graph.
4. We built the Minimum Spanning Tree of each graph.
5. We browsed and analyzed paths in the Minimum Spanning Tree for extraction of the meaning of 公 in the polymorphemic words.

We reviewed Chinese computational morphology and Chinese word sense disambiguation literature and found no prior work proposing such a com-

[1] Actually, this list includes some idioms like 天公不作美 which could not be satisfactorily labeled as polymorphemous words.

[2] source: www.chineseetymology.org/ The list is still not exhaustive!

putational approach for meaning identification of a polysemous morpheme in Chinese words.

We know of no Chinese dictionary or database which lists for each meaning of a polysemous morpheme all the Chinese words embedding the morpheme with this meaning. For example, the Chinese Wordnet of the Academica Sinica[3] proposes a list of some of the different meanings of 公 but provides no listing of all the 公 words with a same given meaning of 公 e.g. *"fair"*.

Our primary research goal is to design tools for Chinese cognitive scientists and linguists who study the semantic interaction between Chinese morphemes and polymorphemic words. Our tools will serve to prepare experimental materials for lexical decision tasks and relatedness judgment tasks involving the repetition of a same Chinese polysemous morpheme embedded with a fixed identified meaning in different Chinese words. [1,2].

2 THE NESTED SEMANTIC LATENT SEMANTIC ANALYSIS SPACES

We used the Academica Sinica Balanced Corpus (ASBC), a five million words corpus based on Chinese materials from Taiwan. The corpus is made of 9183 documents which are considered as semantically meaningful units. Most of the functional words were removed from the corpus.[4]

In the ASBC corpus, 公 as a monomorphemic word occurs with 5 different POS tags: "公(Vh)", "公(Nb)", "公(Nc)", "公(Na)" and "公(A)". These 5 公 words and 187 additional polymorphemic 公 words constitute the list of 192 公 words under study.

2.1 *The First Term-Document Matrix (192 Words 3716 Documents)*

The first and smallest of our term-document matrices was obtained through filtering a whole corpus with a list of 192 公 words. The resulting term-document matrix is made of 192 rows - representing the 192 公 words - and 3716 columns - representing all the documents in which at least one of the 公 words occurs -. The term frequency of each 公 word in each document is stored in that term-document matrix. At that level, we know

[3] http://cwn.ling.sinica.edu.tw/

[4] Words with the following POS tags were removed: Dk Di Caa Cbb Nep Nh P Cab Cba DE I T SHI Neu. For more information about the meaning of the tags, please refer to CKIP Technical Report 95-02/98-04

how the 192 公 words co-occur in the ASBC corpus and we voluntary ignore both the huge number of remaining terms in the corpus and the set of documents in which the 公 words do not occur. This minimalist term-document matrix will serve after Latent Semantic Analysis to create our smallest LSA semantic space. This space is thought to be the worst or poorest representation of the semantic relationships between the 192 公 words.

2.2 *The Second Term-Document Matrix (202 Words 4327 Documents)*

We wanted our second LSA semantic space to contain at least ten words that represent *10 etymological dimensions* of 公. These 10 dimensions words were thought to be able to serve as attractors of semantically similar 公 words and eventually as centroids of 公 clusters. These words could serve later to infer the meaning of 公 in 公 words. We first devised a list of twelve words: (公正, 公平, 公開, 公共, 無私, 貴族, 爵位, 父, 岳父, 雄性, 機關, 機構). These twelve words capture 10 relatively different dimensions of meaning of 公. Both the pairs (貴族, 爵位) and (父, 岳父) are semantically redundant. For example the words 貴族 (noble, nobility) and 爵位 (order of feudal nobility) capture the same meaning of nobility. The word 岳父 (father's in law) is an hyponym of 父 (father), they both capture the fatherhood's relationships meanings of 公. Later we could observe which word in each pairwise behaves as the strongest attractor. In the twelve words list, the first four words are 公 words already present in the first semantic space. Thus to create the second semantic space, we added to the initial list of 192 公 words, the words (貴族, 爵位, 父, 岳父, 雄性, 機關, 機構). We also included the words (國際, 國際性) to attract 公 words referring to international metric units (e.g. 公克 (gram), 公分 (centimeter), 公升 (liter)). After filtering the whole corpus with the new list of 202 words, we obtained a term-document matrix of 202 terms and 4327 documents.

2.3 *The Third Term-Document Matrix (283 Words 6798 Documents)*

To create the third LSA semantic space we added to the precedent list of 202 words:

1. words which are key-words occurring in a Chinese dictionary's definitions of some of the 187 polymorphemic 公 words. For example the definition for åĚňéĞŇ (kilometer) is "量詞。計算長度的單位". The words (量詞, 計算, 長度, 單位) were all added for building the third list of terms.

2. words which share some common morphemes to the multimorphemic words of the initial list. (e.g. 女王 shares the 王 morpheme with 王公)
3. a few words (國家(country), 事物 (thing)) which occur in category labels created by two Taiwanese participants in a pilot study of the subjective sorting of the 187 polymorphemic 公 words.
4. and words[5] which were thought to be potential attractors of certain 公 words (e.g. 動物 (animal) for 公鹿 (male deer), 豬公 (male pig), 公里 (cock) or 七矮人 (seven dwarfs) for 白雪公主 (White-Snow)).

After filtering the whole corpus with a new list of 283 words, we obtained a term-document matrix of 283 terms and 6798 documents. This matrix will serve to compute our biggest micro-semantic space. This third semantic space was thought to be semantically complete and rich enough to embed meaningful semantic relationships between the words its contains. Such a micro-size space could be a better start than a whole corpus semantic space to investigate the different meanings of 公 in 公 words.

2.4 *The Three Weighting Schemes*

To each of our three term-document matrices we applied a total of three weighting schemes:

1. The term-document matrix containing the term frequencies m_{ij} was logarithmised by computing:

$$\log(m_{ij}+1) \tag{1}$$

   as a local weighting scheme. The benefit is to reduce the frequency effect between terms in a same document.
2. As a global weighting scheme, we used the Inverse Document Frequency scheme[3,4]. Every row i - representing the term frequencies of $term_i$ - of the term-document matrix is multiplied by:

$$\log_2\left(\frac{Number\ of\ documents\ in\ the\ corpus}{Numbers\ of\ documents\ in\ which\ the\ term_i\ appears}+1\right) . \tag{2}$$

   Such a weighting scheme gives more weight to words with a global low frequency.

[5] Automatic selection of these words is still to be done. These words were added for testing purposes. They can be removed.

3. At the document level - the columns of the term-document matrix - we also applied a weighting scheme. To reduce the effect of the size difference between documents, we multiplied each column of the term-document matrix by:

$$\log_2\left(\frac{Max\,document\,size}{Document\,size}+1\right) \quad . \tag{3}$$

More weight is given to small documents. This document level weighting scheme is preferred to resizing the corpus's meaning unit from the original entire document to paragraph of a given size. Resizing could result in splitting meaningful units in different documents.

2.5 *Singular Value Decomposition And Reduced SVD*

After applying the three weighting schemes to the term-document matrices, we computed their reduced Singular Value Decomposition (SVD).

Given $U=[u_1,...,u_m]\,\varepsilon\,R^{m*n}$ and $V=[v_1,...,v_n]\,\varepsilon\,R^{n*n}$ two orthogonal matrices, the SVD of a term-document matrix A can be written:

$$A=U\Sigma V^T=\sum_{i=1}^{p}\sigma_i u_i v_i^T \text{ with } \Sigma=diag(\sigma_1,...,\sigma_p)\,\varepsilon\,R^{m*n}\ ,\ p=min\{m,n\} \quad . \tag{4}$$

where $\sigma_1\geq\sigma_2\geq\sigma_p\geq 0$ are the singular values.

For example, for the third term-document matrix, we have $m=283$ and n= 6798.

Thus, the full SVD represents terms and documents in a 283 dimensions space.

After several trials[6], we decided to reduce the dimensionality of the LSA spaces by taking into account only the first one hundred singular values. So for our three term-document matrices we operated a reduced SVD to obtain three 100 dimensions spaces. This can be written:

$$A\simeq A_{100}=U_{100}\Sigma_{100}V_{100}^T \quad . \tag{5}$$

where $\Sigma_{100}=diag(\sigma_1,...,\sigma_{100})$ and $\sigma_1\geq\sigma_2\geq\sigma_{100}>0$ are the 100 first non-zeros singular values.

We termed $A_{192,100}$, $A_{202,100}$ and $A_{283,100}$ the three reduced LSA semantic spaces containing respectivelly 192, 202 and 283 words.

[6] We tried different values, including a dimension equal to the lowest dimension of the term-document matrix -the number of terms- but these choices were discarded while comparing the quality of results described in part 4.

2.6 *Cosine Matrix*

To compare semantic similarity between two words in a LSA space, the cosine measurement of the two vectors v_i, v_n representing the two terms is computed as:

$$\cos(v_i, v_n) = \frac{v_i \cdot v_j}{\|v_i\|\|v_j\|} \quad . \tag{6}$$

For each of our created Chinese LSA space, we computed the whole cosine matrix C of all the terms to measure semantic similarity between words.

$$C = \begin{pmatrix} 1 & \cdots & \cos(v_1, v_j) & \cdots & \cos(v_1, v_m) \\ \vdots & \ddots & & & \\ \cos(v_j, v_1) & & \ddots & & \\ \vdots & & & \ddots & \\ \cos(v_m, v_1) & & & & 1 \end{pmatrix} \quad . \tag{7}$$

C is symmetric due to $\cos(v_i, v_j) = \cos(v_j, v_i)$. We computed the three cosine matrices C_{192}, C_{202} and C_{283} whose dimensions are respectively 192*192, 202*202 and 283*283.

2.7 *Dissimilarity Matrix*

From the cosine matrix C, the dissimilarity matrix D is derived.

$$D = \begin{pmatrix} 1 & \cdots & \cdots & \cdots & 1 \\ \vdots & \ddots & & & \vdots \\ \vdots & & 1 & & \vdots \\ \vdots & & & \ddots & \vdots \\ 1 & \cdots & \cdots & \cdots & 1 \end{pmatrix} - C \quad . \tag{8}$$

$$\text{with } d_{ij} = 1 - \cos(v_i, v_j) \geqq 0$$

We computed the three dissimilarity matrices D_{192}, D_{202} and D_{283} whose dimensions are respectively 192*192, 202*202 and 283*283.

3 GRAPH-THEORY BASED APPROACH

3.1 *A Few Definitions*

A *graph G* = (V,E) is an ordered pair, where V is a set whose elements are called *vertices*, and where E is a set of pairs of distinct vertices. Given *p* and *q* two vertices of V, the element {p,q} ε E is called an *edge* and link the vertices *p* and *q*.

When edges are given a weight - a real number here -, the graph is said to be *weighted*. If no orientation is assigned to edges, the graph is said to be *undirected*. When for every pair of vertices V_i, V_j, there is a sequence of edges allowing to join V_i and V_j, then the graph *G* is said to be connected. If every pair of vertices in *G* is directly connected through an edge, the graph is said to be *complete*. Two vertices V_i and V_j linked by an edge are said to be *adjacent*.

The *adjacency matrix A* of a complete weighted graph *G* is the matrix whose entry A_{ij} is 0 if $i = j$ and otherwise is w_{ij} the weight assigned to the edge V_i, V_j[5,6].

A *tree* of a graph *G* is a connected subgraph of *G* with no cycle. A *spanning tree* (ST) of a graph *G* is a tree of *G* which contains all the vertices of *G*.

A *minimum spanning tree (MST)* of a graph *G* is a spanning tree (ST) of *G* whose the sum of edges is minimum[5,6]. This can be written:

$$\sum_{e \varepsilon MST} w(e) = \min_{ST \varepsilon G} \left(\sum_{e \varepsilon ST} w(e) \right) \quad . \tag{9}$$

3.2 *Applying Graph Theory to the Dissimilarity Matrix*

The dissimilarity matrix *D* introduced in Âğ2.7 can be viewed as the adjacency matrix of a complete weighted undirected graph *G*. The rows and the columns of the adjacency matrices represent the words under study. Each word is a vertex of *G* and each edge of *G* linking two vertices v_i and v_j is weighted by d_{ij}. Thus we have:

$$\forall i \, , \, d_{ii} = 0 \text{ and } \forall (i, j) \text{ with } i \neq j \, , \, d_{ij} = 1 - \cos(v_i, v_j) \quad . \tag{10}$$

From each of the three dissimilarity matrices D_{192}, D_{202} and D_{283}, we used Prim's algorithm to build three minimum spanning trees MST_{192}, MST_{202} and MST_{283} [7]. Hence, each of our LSA semantic space $A_{192,100}$,

$A_{202,100}$ and $A_{283,100}$ has an associated minimum spanning tree. Uniqueness of the MST of a graph G is ensured if each edge of G has a different weight. By removing edges of comparatively high weights in the MST, clusters can be formed [8].

Lemma 1. *[9]*
Any two vertices in a tree are connected through a unique path

Therefore in a MST, the path connecting two vertices is unique. The length of the path between two vertices could be measured by:

1. summing the weights of all the edges composing the path.
2. combining the precedent sum with the total number of intermediary nodes.
3. qualitatively summing the number of concepts composing the path.

Length can serve as an indicator of similarity between two words. This similarity can be interpreted as semantic, situational or of other nature. The shorter the length of the path between two words, the closer is their similarity relationship.

We studied the paths from any of the åĚň words to the twelve words representing the etymological dimensions of åĚň. We also looked at the paths from the twelve dimensions words to the five åĚň morphemes with different POS tags.

4 RESULTS

4.1 *Uniqueness of the Three MST*

For each of the three adjacency matrices D_{192}, D_{202} and D_{283}, some edges have a same weight. Therefore, we concluded than none of our three minimum spanning trees MST_{192}, MST_{202} and MST_{283} is unique.

4.2 *A 192 Vertices MST MST_{192}*

The first MST contained only all the 亼 words.

For Chinese native readers, few of the 191 edges of the MST_{192} between polymorphemic 亼 words bear relevant semantic similarity information. We listed some examples of such edges in Table 1.

Table 1. Edges capturing genuine semantic similarity

Edge	Weight
公秉(nf) – 公升(nf)	4.675626e-03
有限公司(nc) – 公司法(na)	1.304338e-02
公民權(na) – 公民(na)	3.671904e-01
公費生(na) – 公費(na)	9.534522e-02
豬公(na) – 大豬公(na)	8.965163e-06
蘇花公路(nc) – 橫貫公路(nc)	5.542596e-03

MST_{192} captures some hyponomic relationships: 公秉 (kiloliter) and 公升 (liter), or situational relationships: 聖誕老公公 (Father Christmas) and 百貨公司 (department store) as Father Christmas can be found in department store around Christmas.

4.3 *A 202 Vertices MST* MST_{202}

MST_{202} embeds all the 公 words and the selected words representative of dimensions of meanings of 公. In MST_{202}, on average, words belongs to two edges. The twelve dimensions words, on average, also share two edges with other words. Of all the dimensions words, only 父 and 爵位 serve as hypothesized as strong 公 attractors by attracting respectively 4 and 5 words. For a Chinese reader, there are no genuine semantic relationships between 爵位 and the words forming edges with it. 貴族 - representing the same meaning dimension as 爵位 - behaves as a weak attractor by sharing only one edge with a 公 word. 國際 failed to attract international metric units.

The hyponomic relationship in Table 2 between 岳父 and 父 is captured by one edge between the two words.

Table 2. Edge capturing the "father's in law"–"father" hyponomic relationship

Edge	Weight
父(na) – 岳父(na)	8.157458e-02

Except that the five 公 monomorphemic words are outliers, clustering does no provide additional insightful information than simple browsing of the MST.

4.4 *A 283 Vertices Tree LSA 100 dimensions MST_{283}*

In MST_{283}, on average, words belongs to 2 edges as for MST_{202}. The average for the dimensions words is in increase, slightly over 2 (e.g. 2.16). Compared to the two precedent MST, MST_{283} can be used to extract genuinely the meaning of a 公 in a 公 word.

INFERRING THE MEANING OF 公 IN 公鹿 (MALE DEER). Table 3 lists the three edges forming the path from 公鹿 (male deer) to 雄性 (male or maleness) and one edge joining 雄性 and one of the monomorphemic 公 word 公(A).

Table 3. Sequence of Edge capturing the "maleness" meaning of 公 in 公鹿 (male deer)

Edge	Weight
雌(a) – 公鹿(na)	2.363839e-03
雌性(na) – 雌(a)	6.576066e-03
雌性(na) – 雄性(na)	6.728410e-02
雄性(na) – 公(a)	1.832874e-01

Figure 1 represents graphically these four edges. The morpheme 公(A) also shares two additional edges with two polymorphemic 公 words - 公Word in Fig. 1 -.

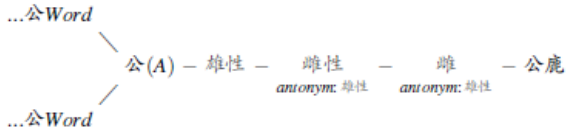

Fig. 1. Paths from 公鹿 (male deer) to 雄性 (maleness) and from 公(A) to 雄性

The two intermediary words 雌性 (female or femaleness) 雌 (female) between 公鹿 and 雄性 are non-公 words and are both antonyms to 雄性. We can say that the path from 公鹿 to 雄性 is conceptually of length 1: only one concept (femaleness) separates the concept of 公鹿 and 雄性.

Besides, 雄性 which is one of our dimension word has attracted a monomorphemic 公 word 公(A). This can mean that one of the meaning of 公(A) is related to 雄性. 公(A) shares two other edges with the words 外公 and 公使. All the three edges are listed in table 4.

Table 4. The three edges of 公(A)

Edge	Weight
公使(na) – 公(a)	6.327506e-01
外公(na) – 公(a)	9.058475e-01
雄性(na) – 公(a)	1.832874e-01

The edge {雄性, 公} has the smallest weight. Thus we can attach 雄性 as a primary meaning to 公.

From the three propositions:

1. In all MST_{283}, 公鹿 shares only one edge with a word: 雌
2. Only one concept (femaleness) separates the concept of 公鹿 and 雄性.
3. 雄性 is a primary meaning of 公(A).

We can infer that in MST_{283}, the closest meaning of 公 in 公鹿 (male deer) is 雄性 (male, maleness). Every Chinese speaker will agree on the meaningfulness and correctness of such a conclusion.

VISUAL REPRESENTATION OF MST_{283} AND CLUSTERING.

MST_{283} is plotted on Fig.2. åĚň dimension words and monomorphemic words are represented with bigger circles to ease their localisation in the MST. The MST_{283} contained the paths between any pairs of words. By removing some of the edges of MST_{283}, clusters[7] can be formed. For example, the five word {公鹿, 雄性, 雌性, 雌, 公(A)} of the example detailed in Âğ 4.4.1 constitute one of the clusters. The mean size of the 30 clusters is 3 and 190 out of 283 words were classified as outliers.

Actually clusters to be efficiently used for meaning extraction, should be represented as subgraphs and not just as sets of words. In the latter case, clustering results are an impoverished representation of the whole knowledge embedded in the minimum spanning tree. The main reason is that the path structure - sequence of vertices to go from one word to another - is not present in clusters. However, considering the cluster {公鹿, 雄性, 雌性, 雌, 公(A)}, it is still possible to infer that the meaning of 公 in 公鹿 is represented by a common conceptual meaning of the three words (雄, 雌性, 雌).

[7] [10,8] showed that clustering from the minimum spanning tree is equivalent to single-linkage clustering.

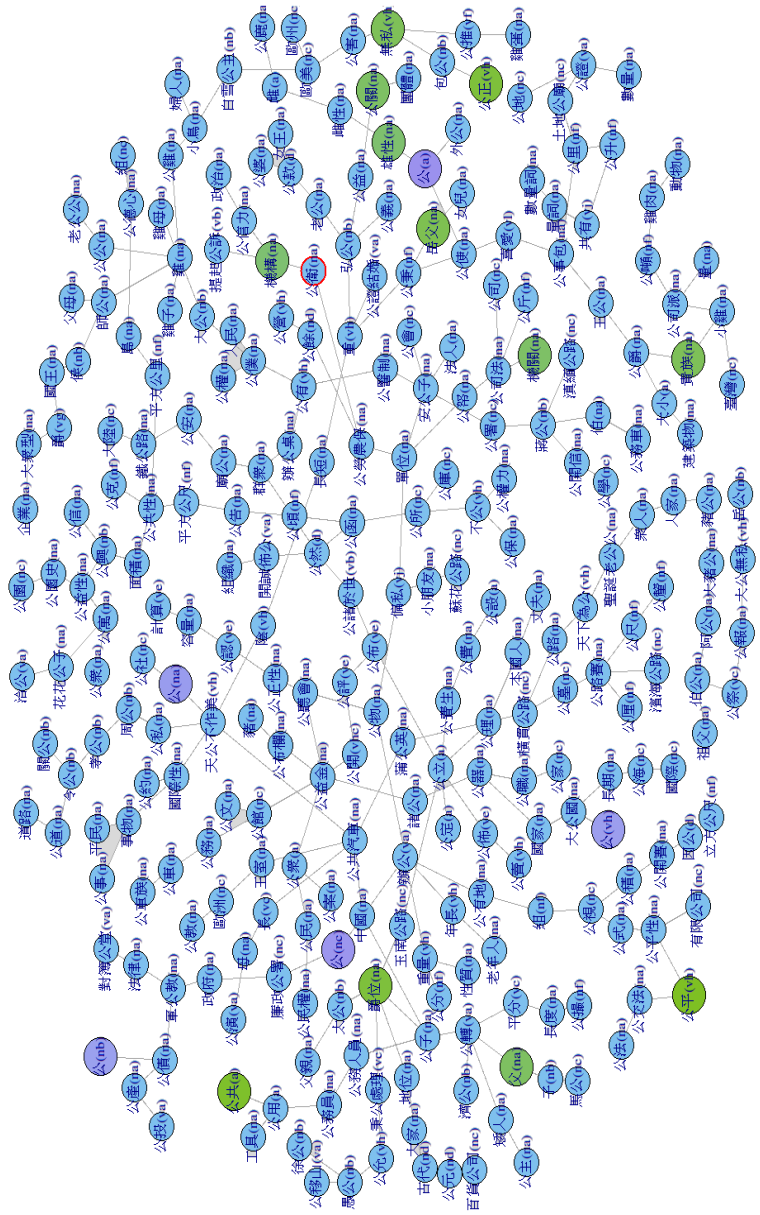

Fig. 1. MST_{283}

5 General Conclusion

Of the three minimum spanning trees, only the biggest - the one which embeds words from the dictionary's definition of the 人 words - can capture the meaning of 人 in 人 polymorphemic words in a way that is satisfactory for a native Chinese reader. In addition to capturing what appears for the observer to be semantic relationships, the edges of the minimum spanning trees can also embed situational relationships.

Finally, it is concluded that our approach is a first step in modeling partly representation and acquisition of the different meanings of Chinese polysemous morphemes. This work is thought to bring some insights to the Plato's problem and additional evidence towards the plausibility of words serving as ungrounded symbols[11,3]. More practically, this work could serve to add a new feature to current Chinese Wordnets: the listing of all the Chinese words embedding a same polysemous morpheme with a fixed identified meaning. Such a listing will help cognitive scientists studying the effects of repetitive exposure to Chinese polysemous morphemes embedded in compound words.

6 Future Directions

Firstly, we aimed at replicating that work using the Chinese Wikipedia instead of the Academica Sinica Balanced Corpus. The Chinese Wikipedia could reflect more adequately the representation of human knowledge as it has a semantic organization and its content and files structure follow categorization meaningful to human.

Secondly, we are presently investigating how to build minimum spanning trees satisfying constraints. For example, we aim at selectively build a MST which would warranty that a maximum of attractors words share edges with a 人 monomorphemic word and with a maximum of 人 words. Such a MST will serve to extract the meaning of a maximum of 人 words.

Finally, instead of using Latent Semantic Analysis to create the nested semantic spaces, we could use the following alternatives:

1. Fiedlar retrieval: [12] proposed that by considering the term-document matrix as a bipartite graph between the set of words and the set of documents, computing a set of the smallest eigenvalues of the Laplacian matrix of the bipartite graph, one can perform an enhanced kind of LSA analysis where unlikely to traditional LSA, documents and terms are considered equivalent and cohabiting in a same space.

2. Probabilistic models of semantic analysis: Latent Dirichlet Allocation (LDA) or Probabilistic LSA. They are probabilistic successors of LSA which have been found to outperform LSA[13,14,15] .

7 Acknowledgments

We thanked Iris Huang for her suggestions about the functional words to remove from the corpus and fruitful discussions and Train Min Chen for sharing subjects' data of her pilot study of a categorization task of 大 words. We also thanked Ingo Feinerer and Fridolin Wild from the R project, for their help in fixing some problems while we were creating and experimenting with our Chinese LSA spaces.

All of the data presented in this paper is freely available from the first author.

References

1. Chen, J.Y., Galmar, B., Su, H.J.: Semantic satiation of chinese characters in a continuous lexical decision task. In: The 21st Annual Convention of the Association For Psychological Science. (2009)
2. Galmar, B., Chen, J.Y.: Can neural adaptation occur at the semantic level? a study of semantic satiation. In: The 12th annual meeting of the Association for the Scientific Study of Consciousnesss. (2007)
3. Landauer, T., Dumais, S.: A solution to plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. Psychological review **104**(2) (1997) 211–240
4. Landauer, T., McNamara, D., Dennis, S., Kintsch, W.: Handbook of latent semantic analysis. Lawrence Erlbaum (2007)
5. Foulds, L.: Graph theory applications. Springer (1995)
6. Gross, J., Yellen, J.: Graph theory and its applications. CRC press (2006)
7. Graham, R., Hell, P.: On the history of the minimum spanning tree problem. Annals of the History of Computing **7**(1) (1985) 43–57
8. Zahn, C.: Graph-theoretical methods for detecting and describing gestalt clusters. IEEE Transactions on computers **20**(1) (1971) 68–86
9. Wu, B., Chao, K.: Spanning trees and optimization problems. Chapman & Hall (2004)
10. Gower, J., Ross, G.: Minimum spanning trees and single linkage cluster analysis. Journal of the Royal Statistical Society. Series C (Applied Statistics) **18**(1) (1969) 54–64
11. Glenberg, A., Robertson, D.: Symbol grounding and meaning: A comparison of high-dimensional and embodied theories of meaning. Journal of Memory and Language **43**(3) (2000) 379–401

12. Hendrickson, B.: Latent semantic analysis and Fiedler retrieval. Linear Algebra and its Applications **421**(2-3) (2007) 345–355
13. Blei, D., Ng, A., Jordan, M.: Latent dirichlet allocation. The Journal of Machine Learning Research **3** (2003) 993–1022
14. Blei, D., Griffiths, T., Jordan, M., Tenenbaum, J.: Hierarchical topic models and the nested Chinese restaurant process. Advances in neural information processing systems **16** (2004) 106
15. Blei, D., Lafferty, J.: A correlated topic model of science. Annals of Applied Statistics **1**(1) (2007) 17–35

BRUNO GALMAR
NATIONAL CHENG KUNG UNIVERSITY,
TAIWAN
E-MAIL: <HSUYUESHAN@GMAIL.COM>

JENN-YEU CHEN
NATIONAL CHENG KUNG UNIVERSITY,
TAIWAN
E-MAIL: <PSYJYC@MAIL.NCKU.EDU.TW>