PBML

The Prague Bulletin of Mathematical Linguistics
NUMBER 94 SEPTEMBER 2010 15–34

Identifying Different Meanings of a Chinese Morpheme through Semantic Pattern Matching in Augmented Minimum Spanning Trees

Bruno Galmar[a], Jenn-Yeu Chen[b]

[a] National Cheng Kung University, Institute of Education
[b] National Cheng Kung University, Institute of Cognitive Science

Abstract

Galmar and Chen (2010) introduced the first corpus-based computational approach to the problem of identifying the different meanings of a polysemous Chinese morpheme embedded in polymorphemic words. The approach is based on the successive application of a dimensionality reduction method - Latent Semantic Analysis, a graph-theory algorithm - Prim's algorithm and a semantic pattern recognition search used for meaning inference. Our present work adds major changes and contributions to Galmar and Chen (2010). Firstly, we theoretically defined and detailed what are the Chinese semantic patterns to be searched in the augmented minimum spanning trees. Then, we modified the computational approach to include the use of Nearest Neighbors lists. This change allows for a major contribution: a proof is given that all the possible outputs of Prim's algorithm in our experiment - the minimum spanning trees - contain all the same amount of semantic information to be used for meaning inference. Thus, our final meaning inference results are optimal.

Practically, this work could serve as a first step to add a new feature to current Chinese Wordnets: the listing of all the Chinese words embedding a certain polysemous morpheme with a fixed identified meaning. Finally, future directions are sketched.

KEYWORDS: Chinese Polysemous Morphemes, Meaning Inference, Dimensionality Reduction, Graph-Theory Analysis, Semantic Patterns.

Corresponding author: hsuyueshan@gmail.com
Cite as: Bruno Galmar, Jenn-Yeu Chen. Identifying Different Meanings of a Chinese Morpheme through Semantic Pattern Matching in Augmented Minimum Spanning Trees. The Prague Bulletin of Mathematical Linguistics No. 94, 2010, pp. 15–34. doi: 10.2478/v10108-010-0018-2.

1. Introduction

The character 公 is a Chinese polysemous morpheme with more than sixteen dimensions of meaning according to an etymological dictionary[1]. The Chinese words 公鹿 (male deer), 外公 (maternal grandfather) and 公平 (fair) are polymorphemic Chinese words that all embed 公 as a common morpheme. However, for a Chinese native speaker, the meaning of 公 in each of the three words is different:

1. In 公鹿 (male deer), the meaning of 公 is male.
2. In 外公 (maternal grandfather), the meaning of 公 is grandfather.
3. In 公平 (fair), the meaning of 公 is fair.

In the three above examples, the meaning of 公 is defined each time by only one word: "male", "grandfather" and "fair". These one-word definitions have a psychological reality for the Chinese native speaker. An interesting fact is that these one-word definitions exist as the etymological dimensions of meaning of 公 [2] given in an etymology dictionary of Chinese. Each etymological dimensions of meaning can be represented by a Chinese word – which we call a meaning dimension word. Thereafter, we retain the linguistically simplified definition of the meaning of 公 in a polymorphemic word as being one meaning dimension word.

In (Galmar and Chen, 2010) and in the present study, the proposed computational approaches aim at identifying the meaning of 公 in a polymorphemic 公 word using knowledge of the etymological dimensions of meaning for the Chinese monomorphemic word 公. Basically, the approaches consist in detecting in a data structure - here a minimum spanning tree - some semantic patterns which are instances of predefined abstract semantic patterns. These patterns can be interpreted through predefined meaning inference rules to identify which meaning dimension word can serve as the meaning of 公 in a polymorphemic 公 word captured in a semantic pattern. So, both (Galmar and Chen, 2010) approach and ours can be viewed as light knowledge-based approaches. We used the adjective "light" to emphasize that the amount of knowledge fed to the meaning inference system is kept as low as possible: a dozen of meaning dimensions words and a few meaning inference rules. Our approach can also discover other potential meaning dimensions words and can eventually update the initial list of meaning dimensions words. Galmar and Chen (2010)'s work and our present study are both generalizable to all Chinese polysemous monomorphemic words. Galmar and Chen (2010) focused on 公 polymorphemic words because experimental data from a categorization label task experiment are available for the 公 words.

[1] source: `http://www.chineseetymology.org/`www.chineseetymology.org/ Some dimensions are overlapping.

[2] The dimensions of meaning for 公 are: 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

Galmar and Chen (2010) observed that there is currently 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] offers 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"*.

Galmar and Chen (2010) reviewed the literature on Chinese computational morphology and Chinese word sense disambiguation. They found no prior work proposing a computational approach to the task of meaning identification of a polysemous morpheme embedded in Chinese words. Therefore, Galmar and Chen (2010) proposed the first corpus-based computational approach to this task. In an example, they showed how the approach can led to extract correctly the meaning of 公 in 公鹿 (Galmar and Chen, 2010).

Galmar and Chen (2010)'s work suffers from the following main limitations:

1. The authors found that there was not a unique minimum spanning tree to serve as a solution. However, only one solution was presented without indication about its selection. The authors gave neither information about possible criteria to use to select the best solution, neither a proof that the presented solution was the best one.
2. The concept of augmented minimum spanning trees did appear briefly but was not detailed and emphasized.
3. Some primary considerations for future directions were absent.

The present work adds the following major contributions and changes to (Galmar and Chen, 2010):

1. The novel computational approach has been designed in a way that allows to prove its optimality. Now, computing the Nearest Neighbors lists plays a major role in helping to prove optimality of the results. We described the novelties of our approach carefully.
2. A new way of formalizing Chinese free morphemes and words interaction for semantic inference is presented for the first time through the concepts of semantic pattern matching and semantically augmented minimum spanning trees.
3. Proofs that our results are optimal are given using solid reasoning based on the work of Prim in graph-theory (Prim, 1957).
4. New detailed examples of meaning inference for a Chinese polysemous morpheme are given to illustrate the results.
5. Future enhancements, directions and generalizations of this work are given.

This work is intended to serve as a first step in:

1. Designing tools for Chinese cognitive scientists and linguists who study the semantic interaction between Chinese morphemes and polymorphemic words. It can help in preparing experimental materials for lexical decision tasks and relatedness judgment tasks involving the repetition of a same Chinese polyse-

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

mous morpheme embedded with a fixed identified meaning in different Chinese words (Chen et al., 2009; Galmar and Chen, 2007).
2. Enhancing Chinese Wordnets with the listing for each meaning of a polysemous morpheme of all the Chinese words embedding the morpheme with this meaning.

The new computational approach to infer the meaning of 公 in polymorphemic words can be unfolded in six steps:

1. The first step encompasses the creation of semantic spaces and their dimensionality reduction. We built three nested lists of words. The Academica Sinica Balanced Corpus (ASBC) was then filtered by these three lists to output three term-document matrices (TDM). The matrices were weighted and then their dimensionality was reduced through the computation of the reduced Singular Value Decomposition (SVD). The final matrices represent the three nested Latent Semantic Analysis (LSA) semantic spaces.
2. The second step consists in computing for each LSA semantic space the cosine matrix and the dissimilarity matrix for all terms.
3. In the third step, each dissimilarity matrix is viewed as the adjacency matrix of a complete weighted undirected graph and is used to build a minimum spanning tree (MST) by applying Prim's algorithm.
4. Besides building the MST, a nearest neighbors (NN) list is built for each graph. The NN list serves to augment the MST with neighboring information.
5. In the fifth step, it is proven that even if the outputed MSTs are not unique, they embed the maximum amount of information useful for meaning inference.
6. In the last step, the paths of the MSTs are browsed in search of patterns to extract the meaning of 公 in the polymorphemic 公 words.

The remainder of this paper is organized as follows. In Section 2, firstly, we newly put forward which particular features of the Chinese language and the ASBC corpus are at the heart of the working of the computational approach. Then, we present briefly Steps 1 and 2 which are fully described in (Galmar and Chen, 2010). Section 3 encompasses Steps 3, 4 and 5. For Step 3, we emphasize the rationale of reducing the completeness of the dissimilarity matrix to obtain a MST. Step 4 and 5 are new steps. In Section 3, we also define and list the semantic patterns to be used for meaning inference. In section 4, the application of Step 6 is illustrated through new results. Then come the conclusion and the future directions sections.

2. The Nested Semantic Latent Semantic Analysis Spaces and Their Dimensionality Reduction

2.1. Philosophy of the Computational Approach and Creation of the Semantic Spaces

At the heart of the design of the present computational approach are the following pecularities of the Chinese language and the ASBC corpus:

1. Some Chinese characters are free morphemes: they are not only embedded in compound words but can also be standalone words (Packard, 2000). 公 is such a free morpheme and is also polysemous.
2. In the ASBC corpus, 公 occurs as a stand-alone monomorphemic word with more than one part-of-speech (POS) tag: it has five different POS tags[4].

The plurality of the occurence of 公 with different POS tags is thought to help the capture of the meaning of 公 in 公 polymorphemic words. We imagined that the 公 monomorphemic words could be captured in a structure - e.g. a graph - with other polymorphemic 公 words forming together semantic patterns. These semantic patterns could be used to infer some different meanings of the polysemous 公 morpheme.

To employ a metaphor inspired by the software hacking culture, we viewed polysemous free morphemes like 公 as a potential backdoor to the semantics of Chinese polymorphemic words. Hence, the work is an attempt to crack the semantics of the Chinese language using some pecularities of the Chinese language and of the annoted ASBC corpus.

There is also a flavor of whole-part thinking in our approach: the part - the free morpheme - and the whole - the compound word - are both necessary to conduct semantic inference about the identification of the meaning of a Chinese polysemous morpheme.

2.2. The Building of Nested Semantic Latent Semantic Analysis Spaces

Galmar and Chen (2010) built three nested lists of 公 words extracted from the 5 million words Academica Sinica Balanced Corpus (ASBC). Full details about how were built these three lists is given in Galmar and Chen (2010). These three lists are used to filter the ASBC corpus to obtain three term-document matrices (TDM). The three TDM will serve to build the three semantic LSA spaces by application of Latent Semantic Analysis.

The main idea is to create three semantic spaces of increasing semantic richness:

1. The smallest semantic space contains only 公 words. This space is thought to be the poorest representation of the semantic relationships between the 192 公 words. The 5 公 words and 187 additional polymorphemic 公 words constitute the initial list of 192 公 words under study. The TDM was made of 192 words and 3716 documents.
2. The second semantic LSA space contains words that represent ten etymological dimensions of the meaning of 公. These meaning dimension words could attract or be attracted by semantically similar 公 words in semantic patterns in a graph structure. From a cluster-based viewpoint, these words could serve as centroids of 公 words clusters. They could also be used later to infer the meaning of 公 in 公 words. The TDM was made of 202 words and 4327 documents. The twelve

[4]" 公 (Vh)", " 公 (Nb)", " 公 (Nc)", " 公 (Na)" and " 公 (A)"

meaning dimension words are: (公正, 公平, 公開, 公共, 無私, 貴族, 爵位, 父, 岳父, 雄性, 機關, 機構)[5].

3. The third and biggest semantic space is thought to be semantically rich enough to embed meaningful semantic relationships between the words it contains. Such a micro-size space could be an alternative start to a whole corpus semantic space to investigate the different meanings of 公 in 公 words. The TDM was made of 283 words and 6798 documents. Among the additional words which were inserted in the third list, there were:
   (a) words which are key-words occurring in a Chinese dictionary's definitions of some of the 187 polymorphemic 公 words.
   (b) non-公 compound words that share some common morphemes with the 187 公 compound words.
   (c) a few words (e.g. 國家 (country)) which have been chosen as category labels by two Taiwanese participants in a pilot study of the subjective sorting of the 187 polymorphemic 公 words.

From a language acquisition standpoint, our approach is thought to bear the two following realistic traits: both the size of the lexicon and its semantic richness increase - as they do during language learning -.

2.2.1. The Three Weighting Schemes

To each of the three term-document matrices, Galmar and Chen (2010) applied a total of three weighting schemes:

1. A local weighting scheme. The TDM containing the term frequencies m_{ij} is logarithmised. Hence, the effect of frequency differences between terms in a same document is reduced.
2. A global weighting scheme. The classic Inverse Document Frequency scheme (Landauer and Dumais, 1997; Landauer et al., 2007) is used. It gives more weight to words with a global low frequency.
3. At the document level - the columns of the term-document matrix - a weighting scheme is applied. To reduce the effect of the size difference between documents, each column of the term-document matrix is multiplied by:

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

   More weight is given to small documents. Such a choice is motivated by the heuristic that especially for news articles - the documents of the ASBC corpus -, it is easier to extract the gist of a short article than a very long one: it means that small documents are generally better informative than long ones about their

[5] In the twelve words list, the first four words are 公 words already present in the first semantic space. These twelve words capture ten relatively different dimensions of meaning of 公. Two of the words capture the same dimension meaning. See (Galmar and Chen, 2010) for more details.

inherent meaning[6]. 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.

2.2.2. Singular Value Decomposition and Reduced SVD

After being weighted, the three TDM have their reduced Singular Value Decomposition (rSVD) computed. Galmar and Chen (2010) explained why they empirically decided to reduce the dimensionality of the LSA spaces by taking into account only the first one hundred singular values. This rSVD can be written:

$$A \simeq A_{100} = U_{100}\Sigma_{100}V_{100}^{T} \quad . \tag{2}$$

where A is the original TDM, A_{100} the reduced TDM, U_{100} and V_{100} two orthogonal matrices[7] and 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.

Galmar and Chen (2010) termed $A_{192,100}$, $A_{202,100}$ and $A_{283,100}$ the three reduced LSA spaces containing respectivelly 192, 202 and 283 words. $A_{192,100}$, $A_{202,100}$ and $A_{283,100}$ are the Chinese LSA spaces that will be used to compute the cosine and dissimilarity matrices.

2.2.3. Cosine Matrix and Dissimilarity matrix

In $A_{192,100}$, $A_{202,100}$ and $A_{283,100}$, words are represented as vectors. Similarity between two words of a Chinese LSA space is measured by calculation of the cosine value between the two corresponding vectors. For each of the three Chinese LSA spaces, Galmar and Chen (2010) built the whole cosine matrix C defined by $c_{ij} = \cos(v_i, v_j)$, where i and j are two words of a Chinese LSA space and where v_i and v_j their representating vectors. C is symmetric due to $\cos(v_i, v_j) = \cos(v_j, v_i)$. The three cosine matrices C_{192}, C_{202} and C_{283} were computed.

From the cosine matrix C, the dissimilarity matrix D is derived with:

$$d_{ij} = 1 - c_{ij} = 1 - \cos(v_i, v_j) \geqq 0 \quad . \tag{3}$$

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

---

[6] An illustrative comparison would be that of politicians' speeches. Politicians are sometimes criticized to drown the meaning of their speech in their length. After listening to a long politician's speech one could have the feeling of having lost its gist. Politicans could not achieve to induce this feeling with a very short speech while attention of listeners is remaining high.

[7] U_{100} and V_{100} are the truncated matrices of the orthogonal matrices occuring in the SVD of A.

3. The Graph-Theory Based Approach

3.1. A Few Definitions

Galmar and Chen (2010) introduced all the basic concepts from graph theory which will be used thereafter. Here, we just define again the most crucial concepts and add two definitions.

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 $\{V_i, V_j\}$ (Gross and Yellen, 2006).

A *nearest neighbor (NN)* V_{NN} of a vertex V_i is a vertex for which the weight of the edge $\{V_{NN}, V_i\}$ is minimum among all the edges joining V_i . A *second nearest neighbor (SNN)* V_{SNN} of a vertex V_i is the second vertex of smallest weight and follows V_{NN} in the NNs list. Two vertices V_i and V_j are said to be *reciprocal nearest neighbors (RNN)* if V_i is the nearest neighbor of V_j and vice versa. In the same way, two vertices V_i and V_j are said to be *secondary reciprocal nearest neighbors (SRNN)* if V_i is the second nearest neighbor of V_j and vice versa.

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 (Foulds, 1995). This can be written:

$$\sum_{e \varepsilon MST} w(\mathrm{e}) = \min_{ST \varepsilon G} \left(\sum_{e \varepsilon ST} \mathrm{w(e)} \right) \quad . \tag{4}$$

In a *short-path tree (SPT)*, the length of the paths between the root vertex and all the other vertices are minimum.

3.2. Applying Graph Theory to the Dissimilarity Matrix

3.2.1. Building the Adjacency Matrix.

Galmar and Chen (2010) viewed the dissimilarity matrix D defined in §2.2.3 as the adjacency matrix of a complete weighted undirected graph G. Each word of a given Chinese LSA space 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:

$$\forall\, i\ ,\ d_{ii} = 0 \text{ and } \forall\, (i,j) \text{ with } i \neq j\ ,\ d_{ij} = 1 - \cos(v_i, v_j) \quad . \tag{5}$$

3.2.2. Building the Nearest Neighbors Lists.

From each of the dissimilarity matrices D_{192}, D_{202} and D_{283}, we computed the nearest neighbors lists NN_{192}, NN_{202} and NN_{283} . The i-th member of the list NN_X is the NN of the i-th member of the original list of X 公 words.

3.2.3. Building One Minimum Spanning Tree.

From D_{192}, D_{202} and D_{283} we can directly build three minimum spanning trees MST_{192}, MST_{202} and MST_{283} using Prim's algorithm (Prim, 1957; Graham and Hell, 1985). Hence, each of our semantic LSA space $A_{192,100}$, $A_{202,100}$ and $A_{283,100}$ has an associated minimum spanning tree.

In (Prim, 1957), Prim enunciated two construction principles for building a MST of an undirected weighted graph G:

1. Any unconnected vertex can be directly connected to a nearest neighbor (P1).
2. Any unconnected sub-MST[8] can be connected to a nearest neighbor by a shortest available path (P2).

Prim went on by validating these two construction principles by proving the two following necessary conditions (NC1 and NC2) for a MST:

1. Every vertex in a MST is directly connected to at least one nearest neighbor (NC1).
2. Every subgraph of a MST is connected to at least one nearest neighbor by a shortest available path (NC2).

In Prim's algorithm, P1 is first applied to build a first edge and then P2 is continuously applied to grow from the very first edge a subgraph that will be a MST once all the vertices of the initial graph G will have been connected.

3.2.4. Uniqueness of the Minimum Spanning Tree.

Uniqueness of the MST of a graph G is ensured if each edge of G has a different weight (Eppstein, 1995). In case of non-uniqueness of the MST, solutions provided by (Eppstein, 1992, 1995; Broder and Mayr, 1997; Wright, 1997, 2000) can be applied to count and enumerate all the MSTs.

3.2.5. Counting and Enumerating All the MSTs.

(Eppstein, 1995) proposed to build an equivalent graph EG of the graph G such that the ST of EG are the MSTs of G. The *Kirchoff's matrix-tree* theorem (de Abreu, 2007) served to compute the number of ST in EG corresponding to the number of MSTs in G. Listing all the STs of EG gives all the MSTs of G. (Wright, 1997, 2000) proposed another procedure to count the number of MSTs and build them.

3.2.6. Patterns to Be Observed in MSTs for Meaning Inference.

Once a MST is obtained for the semantic space $A_{X_{words},100}$, it will be used for meaning inference. Here, we are concerned by describing the expected patterns of vertices - to be found in the MST - that will serve for meaning inference. Each pattern is

[8] In the original paper, Prim used *isolated fragment* to refer to a subgraph of a MST. Here we use the word sub-MST.

paired with a set of assertions concerning the meaning inferences that can be formulated. Figure 1 presents such an ideal pattern. In Fig. 1, one meaning dimension word (*DimWord*) - introduced in §2.2 - shares an edge with a 公 monomorphemic word $公_{\text{mono}}$. The latter shares one or more edges with 公 polymorphemic words. The weight of the edge {$公_{\text{mono}}$, *DimWord*} is assumed to be the smallest one among the edges joining $公_{\text{mono}}$. In other words, *DimWord* is the NN of $公_{\text{mono}}$. This is represented on Fig.1 by a left-to-right arrow from $公_{\text{mono}}$ to *DimWord* labeled with the symbol NN. In the semantic space $A_{X_{words},100}$, for the pattern represented by Fig.1, we will assert that the primary meaning of $公_{\text{mono}}$ is the meaning of *DimWord*. $公_{\text{Word target}}$ is a 公 polymorphemic word directly connected to *DimWord* and shares no other edge with other words. Therefore, in $A_{X_{words},100}$, the meaning of 公 in $公_{\text{Word target}}$ is the meaning of $公_{\text{mono}}$ whose primary meaning is the meaning of *DimWord*.

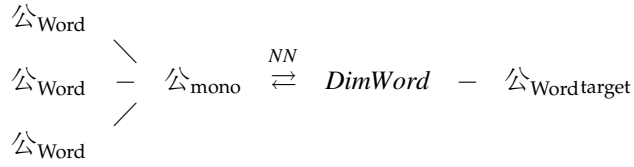

Figure 1. An ideal component of a MST for meaning inference.

The pattern presented in Fig. 2 is an extension of Fig. 1 where DimWord and $公_{\text{Word target}}$ are connected through a number N of non-公 Word. In that case, we will also assert that in $A_{X_{words},100}$, the closest meaning to the meaning of 公 in $公_{\text{Word target}}$ is the meaning of $公_{\text{mono}}$ which itself has for meaning the meaning of *DimWord*.

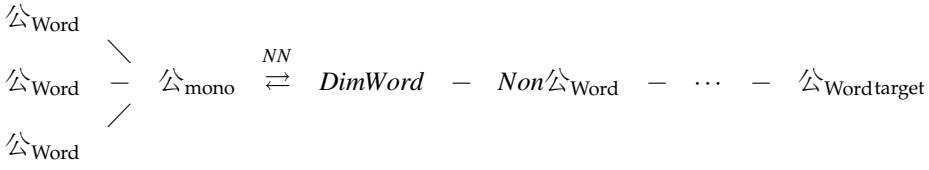

Figure 2. Another ideal component of a MST for meaning inference

In Fig. 3, *DimWord* and $公_{\text{mono}}$ are reciprocal nearest neighbors (RNN). We will assert that the meaning of $公_{\text{mono}}$ in $A_{X_{words},100}$ is the meaning of *DimWord*. Of all the edges joining respectively $公_{\text{Word target1}}$ and $公_{\text{Word target2}}$, the edges {*DimWord*, $公_{\text{Word target1}}$}

and {公$_{\text{mono}}$, 公$_{\text{Word target2}}$} are assumed to be of smallest weight. This means that 公$_{\text{Word target1}}$ and 公$_{\text{Word target2}}$ have respectively *DimWord* and 公$_{\text{mono}}$ as NN. In this case, we will assert that in $A_{X_{words},100}$, the meaning of 公 in 公$_{\text{Word target1}}$ and in 公$_{\text{Word target2}}$ is the meaning of 公$_{\text{mono}}$ which itself has for meaning the meaning of *DimWord*.

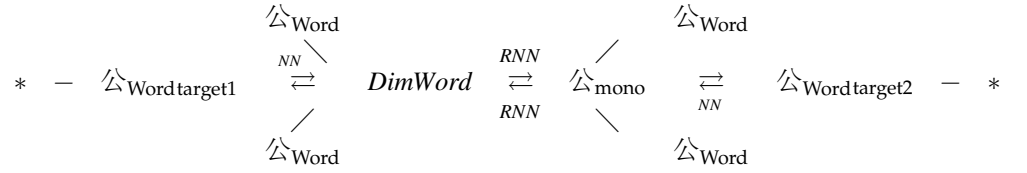

Figure 3. A more complex ideal component

The above list of presented patterns for meaning inference is not exhaustive. It was given to illustrate the logic of how meaning inference is conducted from sequences of paths in the MST.

3.2.7. Walking the MST.

Lemma 1. *(Wu and Chao, 2004)*
Any two vertices in a tree are connected through a unique path.

Following Lemma 1, 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. This requires human judgment.
4. quantitatively summing the number of concepts composing the path. This requires to feed the meaning inferring program with some additional knowledge about the relationships between the non-公 words added into the third semantic space. Some of the non-公 words are synonyms, antonyms or hyponyms of other non-公 words. The MST does not embed originally such knowledge. Therefore the MST is to be augmented. That can be easily done by augmenting the structure of vertices in the MST for non-公 words.

The last solution for path length calculation was chosen in this study. Path length can serve as an indicator of similarity between two words. This similarity will be interpreted primarily as semantic but could be situational or of other nature. The shorter the length of the path between two words, the closer is their similarity relationship.

Table 1. Nearest neighbors for the monomorphemic 公 words

Vertex	Nearest Neighbor	Weight
公 (a)	雄性 (na)	0.1832
公 (na)	公社 (nc)	0.3227
公 (nb)	公債 (na)	0.8319
公 (nc)	廉政公署 (nc)	0.4776
公 (vh)	大公國 (na)	0.2364

3.3. Rationale for Applying Graph Theory to the Dissimilarity Matrix

MSTs are thought to be the smallest data structures which connect all the words of a semantic space through paths and which embed core relationships between words. Comparisons of weight values for the different edges joining a given vertex allow to determine all kinds of nearest neighbor relationships (NN, RNN) that have been precedently showed as necessary for meaning inference.

MSTs are considered more general and balanced in information than shortest-path trees (SPT).

4. Results

4.1. Nearest Neighbors

From the three dissimilarity matrices D_{192}, D_{202} and D_{283}, we computed the nearest neighbors list of each vertex. We present only the NNs list for our biggest semantic space $A_{283,100}$.

4.1.1. The Nearest Neighbors NN_{283} of $A_{283,100}$.

Uniqueness of Nearest Neighbors. Of the 283 words, three (周公 (nb), 關公 (nb) and 雞 (na)) have two nearest neighbors. The remaining words have a single - uniquely defined - nearest neighbor.

Nearest neighbors for the monomorphemic 公 words. Table 1 shows the nearest neighbors for the monomorphemic 公 words and the associate weights. The smaller the weight value is, the closer is the relationship between the 公 word and its NN.

The nearest neighbor of 公 can be interpreted as its primary meaning in the semantic space $A_{283,100}$. At best, the NN is one of the hypothesized meaning dimension word. For example, 公 (a) has 雄性 as a nearest neighbour which is also a meaning dimension word. It can be concluded that 雄性 is the primary meaning of 公 (a). Others monomorphemic 公 words do not have a dimension word as a NN.

Table 2. The monomorphemic 公 words as nearest neighbors

Vertex	公 Nearest Neighbor	Weight
外公 (na)	公 (a)	0.9058
公社 (nc)	公 (na)	0.3227
∅	公 (nb)	∅
廉政公署 (nc)	公 (nc)	0.4776
大公國 (na)	公 (vh)	0.2364

Table 3. Nearest neighbors of the 公 meaning dimension words

Vertex	Nearest Neighbor	Weight
公正 (vh)	包公 (nb)	0.0606
公平 (vh)	公交法 (na)	0.0327
公開 (vhc)	公益金 (na)	0.0173
公共 (a)	公用 (a)	0.5621
無私 (vh)	公害 (na)	0.2964
貴族 (na)	公爵 (na)	0.0768
爵位 (na)	地位 (na)	0.0424
雄性 (na)	雌性 (na)	0.0672
父 (na)	子 (nb)	0.3519
岳父 (na)	公秉 (nf)	0.2454
機構 (na)	公衛 (na)	0.1886
機關 (na)	公司法 (na)	0.6194

The monomorphemic 公 words as nearest neighbors. Monomorphemic 公 words attracted 公 words and even became their NNs. No word has 公 (nb) as a NN.

From Tab. 1 and Tab. 2, we noticed the existence of *reciprocal nearest neighbors* relationships (RNN). The NN of 公 (na) is 公社 (nc) and vice-versa. {公 (nc), 廉政公署 (nc)} are RNN and {公 (vh), 大公國 (na)} too.

Nearest neighbors of the 12 meaning dimension words. Table 3 presents the nearest neighbors of the 12 meaning dimension words. Except for 雄性, 爵位 and 父, the dimension words have 公 words as NN.

Table 4. Frequencies of the 公 meaning dimension words as nearest neighbors

Dimension Word	Frequencies as a NN	Source Nodes	Weight
公正 (vh)	1	包公 (nb)	0.0606
公平 (vh)	2	公交法 (na), 公平性 (na)	0.0327, 0.2287
公開 (vhc)	1	公益金 (na)	0.0173
公共 (a)	0		
無私 (vh)	0		
貴族 (na)	2	公爵 (na), 小雞 (na)	0.0768, 0.2430
爵位 (na)	2	公子 (na), 地位 (na)	0.0907, 0.0424
父 (na)	0		
岳父 (na)	1	女兒 (na)	0.6103
雄性 (na)	1	公 (a)	0.1832
機構 (na)	3	公信力 (na), 公衛 (na)	0.6165, 0.1886
		提起公訴 (vb)	0.2226
機關 (na)	0		

Frequencies of the meaning dimension words as nearest neighbors. Table 4 shows how many times the dimension words have served as a nearest neighbor. From Tab. 3 and Tab. 4, we observed that the meaning dimension word 公正 and the 公 -word 包公 are RNN. Therefore in $A_{28_{3,100}}$, the meaning of 公 in 包公 is 公正 (just, fair). Actually, 包公 is the name of an Ancient China high official who symbolizes justice. Chinese speakers will associate differently the meaning of 公 in 包公 to the noun judge or lord which is conceptually close to the meaning of just, fair.

Table 5 lists some vertices with their NNs. Such a list captures genuine semantic similarity.

Table 5. Nearest neighbors capturing genuine semantic similarity

Vertex	Nearest Neighbor
公雞 (na)	雞 (na)
雞 (na)	雞子 (na)
雞母 (na)	雞子 (na)
雌性 (na)	雌 (a)
阿公 (na)	伯公 (na)

4.2. Uniqueness of the Three MSTs

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} may be unique (Eppstein, 1995).

4.3. Counting the Number of MST Meaningful to our Study

The total number of MSTs can be calculated as described in § 3.2.5 . However, we aim at determining the number of MSTs embedding the maximum of information for meaning inference. The MSTs of interest are the MSTs for which the number of patterns defined in §3.2.6 would be maximum. We are going to prove that in our case any MST is the best one: all MSTs contain the same amount of information for meaning inference.

4.4. Any MST_{283} is optimal for meaning inference.

Let's assume $G = (V, E)$ to be the complete connected weighted graph of the semantic space $A_{283,100}$. We preably identified in §4.1.1 the three words that have more than one NNs. These three words and their NNs are neither dimension words neither 公 monomorphemic words and they account for the non-uniqueness of MST_{283}. Two of the three words share the same two nearest neighbors, so that these three words and their NNs form two subparts of the MST and not three ones. These two subparts - sub1MST and sub2MST - will be distinctly connected through any smallest shortest path to the remaining subpart of the MST - sub3MST -. Sub3MST connects all the vertices not belonging to sub1MST and sub2MST.

Lemma 2. [9] *If every vertex of a graph G has only one nearest neighbor, the MST of G is unique.*

Sub3MST's vertices have each only one NN such that by application of Lemma 2, we deduced the uniqueness of sub3MST. Therefore, sub3MST embeds in a unique configuration all of the twelve meaning dimension words and all of the five monomorphemic 公 words and their associated NNs or RNNs. The closest words to these seventeen words are still in sub3MST. By consequence, the patterns of meaning inference are to be found in sub3MST and will not change in the different MSTs.

4.5. Meaning Inference for a 283 Vertices MST: MST_{283}

MST_{283} can be used to extract genuinely the meaning of a 公 in a 公 word.

[9] This lemma can be proved by using Prim's first necessary condition (NC1) for a MST in a simple proof by contradiction as in (Prim, 1957).

4.5.1. Inferring the Meaning of 公 in 公鹿 (Male Deer).

Figure 4 shows the path from one of the monomorphemic 公 word 公 (A) to 公鹿 (male deer). 公 (A) forms an edge with the meaning dimension word 雄性 (male or maleness). The pattern presented by Fig. 4 follows the pattern of Fig. 2 in §3.2.6. 雄性 is the NN of 公 (A), it represents its primary meaning. In Fig. 4, the augmented structure of the vertices of the non-公 words, 雌性 (female or femaleness) 雌 (female) is shown. These two words are both antonyms to 雄性. Following §3.2.7, path length is measured as 1: only one concept (femaleness) separates the concept of 公鹿 and 雄性.

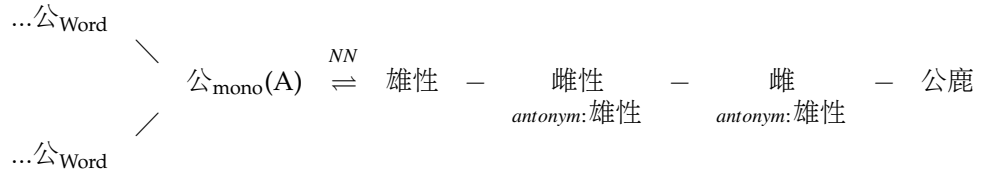

Figure 4. Augmented pattern for inferring the Meaning of 公 *in* 公鹿

As for Fig.2 in §3.2.6, it can be asserted that in $A_{2\mathcal{S}3,100}$, 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.

4.5.2. Inferring the meaning of 公 in 大公國 (Grand Duchy).

From Tab. 1 and Tab. 2, we observed that 公 (vh) and 大公國 are RNN. 大公國 represents an unexpected dimension of meaning of 公. Hence, in $A_{2\mathcal{S}3,100}$, the meaning of 公 (vh) would be the meaning of 大公國. The pattern for meaning inference is presented in Fig. 5. By an examination of edge weights, it is found that 公 (vh) and 國家 (na) are secondary reciprocal nearest neighbors (SRNN).

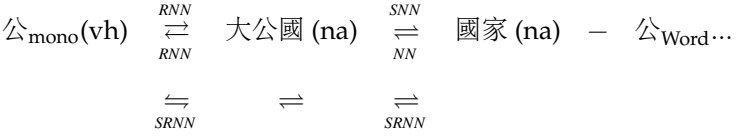

*Figure 5. {*公, 大公國, 國家*} subcomponent of* $MST_{2\mathcal{S}3}$

國家 (country, nation) as mentioned in §2.2, is a possible meaning dimension word for some Taiwanese subjects. If here 國家 is considered as a meaning dimension word, the conclusion is that the meaning of 公 in both 公 (vh) and 大公國 is 國家 (country, nation).

4.5.3. Inferring the meaning of 公 in 廉政公署 (Independent Commission Against Corruption).

In Fig. 6, 公 (nc) and 廉政公署 (nc) are RNN. 廉政公署 (nc) represents an unexpected dimension of 公. The second nearest neighbor of 廉政公署 is 政府 (na). In $A_{28_{3,100}}$, the closest non-公 word to 公 (nc) and 廉政公署 is 政府 (government). Therefore the meaning of 公 in 廉政公署 and 公 (nc) would be 政府. Such an inference is congruent with 公署 in 廉政公署 meaning government office.

$$\text{公}_{\text{mono}}(\text{nc}) \underset{RNN}{\overset{RNN}{\rightleftharpoons}} \text{廉政公署 (nc)} \underset{SNN}{\overset{NN}{\leftrightharpoons}} \text{政府 (na)} \quad - \quad \text{公}_{\text{Word}}\cdots$$

Figure 6. The meaning of 公 *in* 廉政公署 *is* 政府 *(na).*

5. Conclusion

This work brought major changes and improvements to the computational approach introduced by (Galmar and Chen, 2010) for inferring the meaning of a polysemous Chinese morpheme embedded in polymorphemic words. Our modified computational approach uses the similarity matrix of reduced Singular Value Decomposition to build nearest neighbors lists and minimum spanning trees whose vertices structure can be augmented to compute a conceptual distance measure. A constructive proof of optimality of the results have been given.

Our computational approach could serve as a first step 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 could also assist cognitive scientists in preparing materials for experiments aiming at investigating the effects of repetitive exposure to Chinese polysemous morphemes embedded in compound words.

6. Future Directions

Firstly, Galmar and Chen (2010) mentioned that the Academica Sinica Balanced Corpus could not reflect adequately the representation of human knowledge. They advanced the Chinese Wikipedia - with its content organization following categorization meaningful to human - as a sound corpus for a replication of their study.

Secondly, our computational approach can be abstracted and viewed as a computational chain for processing Chinese words semantics. The chain can be unfolded into the following elements:

1. Dimensionality reduction.
2. Completeness reduction. Here the completeness of the graph - whose adjacency matrix is the similarity matrix - is reduced to a minimum spanning tree.
3. Building of an augmented structure with additional semantic relationship information.
4. Search of semantic patterns.
5. Logical Inference.

The first element in the chain, dimensionality reduction was done through Latent Semantic Analysis. Other possibilities would include:

1. Fiedlar retrieval: Hendrickson (2007) proposed that by considering the term-document matrix as a bipartite graph between the set of words and the set of documents and 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. This approach belongs to spectral graph theory (Mohar, 1997; Chung, 1997; de Abreu, 2007).
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 (Blei et al., 2003, 2004; Blei and Lafferty, 2007).

For the second element in the chain, we could:

1. Replace the minimum spanning tree by other kinds of trees such as:
   (a) Light Approximate Shortest-path Tree which is a hybrid tree between the shortest path tree and the minimum spanning tree (Khuller et al., 1995).
   (b) The multi-criteria Minimum Spanning Tree (mc-MST) which takes into account constraints (Zhou and Gen, 1999).
2. Instead of working with the similarity matrix viewed as the adjacency matrix of a complete graph, we could consider to view the similarity matrix as the matrix representation of an hypergraph - a generalization of graphs where edges can join more than 2 vertices (Berge, 1976). Then, we would have to generate a minimum spanning tree for the hypergraph.

A last point, by iterating our approach with the change of the meaning dimensions words and the non-公 words, we could capture the meaning of 公 in more multimorphemic 公 words than in a single iteration.

7. Acknowledgments

We thank Iris Huang and Train Min Chen for fruitful discussions and suggestions concerning the present work.

Bibliography

Berge, C. *Graphs and hypergraphs*. North-Holland Pub. Co., 1976.

Blei, D.M. and J.D. Lafferty. A correlated topic model of science. *Annals of Applied Statistics*, 1 (1):17–35, 2007.

Blei, D.M., A.Y. Ng, and M.I. Jordan. Latent Dirichlet allocation. *The Journal of Machine Learning Research*, 3:993–1022, 2003.

Blei, D., T.L. Griffiths, M.I. Jordan, and J.B. Tenenbaum. Hierarchical topic models and the nested Chinese restaurant process. *Advances in neural information processing systems*, 16:106, 2004.

Broder, A.Z. and E.W. Mayr. Counting minimum weight spanning trees. *Journal of Algorithms*, 24(1):171–176, 1997.

Chen, Jenn-Yeu, Bruno Galmar, and Hsiao-Jen Su. Semantic Satiation of Chinese Characters in a Continuous Lexical Decision Task. In *The 21st Annual Convention of the Association For Psychological Science*, 2009.

Chung, F.R.K. *Spectral graph theory*. Amer Mathematical Society, 1997.

de Abreu, N.M.M. Old and new results on algebraic connectivity of graphs. *Linear Algebra and its Applications*, 423(1):53–73, 2007.

Eppstein, D. Finding the k smallest spanning trees. *BIT Numerical Mathematics*, 32(2):237–248, 1992.

Eppstein, D. Representing all minimum spanning trees with applications to counting and generation. Technical report, Citeseer, 1995.

Foulds, LR. *Graph theory applications*. Springer, 1995.

Galmar, B. and J.Y. Chen. Identifying Different Meanings of a Chinese Morpheme through Latent Semantic Analysis and Minimum Spanning Tree Analysis. *International Journal of Computational Linguistics and Applications*, 1(1-2):153–168, 2010.

Galmar, Bruno and Jenn-Yeu Chen. 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.

Graham, R.L. and P. Hell. On the history of the minimum spanning tree problem. *Annals of the History of Computing*, 7(1):43–57, 1985.

Gross, J.L. and J. Yellen. *Graph theory and its applications*. CRC press, 2006.

Hendrickson, B. Latent semantic analysis and Fiedler retrieval. *Linear Algebra and its Applications*, 421(2-3):345–355, 2007.

Khuller, S., B. Raghavachari, and N. Young. Balancing minimum spanning trees and shortest-path trees. *Algorithmica*, 14(4):305–321, 1995.

Landauer, T.K. and S.T. Dumais. A solution to plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological review*, 104 (2):211–240, 1997.

Landauer, T.K., D.S. McNamara, S. Dennis, and W. Kintsch. *Handbook of latent semantic analysis*. Lawrence Erlbaum, 2007.

Mohar, B. Some applications of Laplace eigenvalues of graphs. *Graph Symmetry: Algebraic Methods and Applications*, 497:227–275, 1997.

Packard, J.L. *The morphology of Chinese: A linguistic and cognitive approach*. Cambridge Univ Pr, 2000.

Prim, R.C. Shortest connection networks and some generalizations. *Bell System Technical Journal*, 36(6):1389–1401, 1957.

Wright, P. Counting and constructing minimal spanning trees. *Bulletin of the Institute of Combinatorics and its Applications*, 21:65–76, 1997.

Wright, P. On Minimum Spanning Trees and Determinants. *Mathematics Magazine*, 73(1):21–28, 2000.

Wu, B.Y. and K.M. Chao. *Spanning trees and optimization problems*. Chapman & Hall, 2004.

Zhou, G. and M. Gen. Genetic algorithm approach on multi-criteria minimum spanning tree problem. *European Journal of Operational Research*, 114(1):141–152, 1999.

Address for correspondence:
Bruno Galmar
hsuyueshan@gmail.com
National Cheng Kung University, Institute of Education
No.1, University Road, Tainan City 701, Taiwan (R.O.C.)
FAX: 886-6-2766493