Structures in CCD

Yu Jiangsheng
©Institute of Computational Linguistics
Peking University

May 14,2001

Topics in CCD

1. lexicon oriented to application

2. SynSet as concept in CCD

3. formal description of structures in CCD

4. dynamic lexicon learning from corpus

Constructing Lexicon

From the viewpoint of Computational Lexicology, a well-structured lexicon should be

1. oriented to some application

2. mathematically formalized

3. computable with limited complexity*

*Referring to automatic construction of lexicon besides the checking of both lexicographic structures and knowledge representation of language.

CCD Construction

Chinese Concepts Dictionary (CCD) is WordNet like semantic lexicon of contemporary Chinese oriented to Information Extraction (IE) and Information Retrieval (IR). The main structure (hypernymy relation) of CCD is described by labeled tree, on which there is a close conceptual reasoning.*

*The structure generates the formation of Chinese phrases from the viewpoint of semantics. See CCD Specification, the further work.

Concept Type in CCD

Definition 1 Let $\Sigma \neq \emptyset$ be the finite set of words in language $\mathcal{L}$, theoretically, $A \in 2^{\Sigma}$ is called a concept. It is easy to see that there are at most $2^{|\Sigma|}$ concepts in $\mathcal{L}$.

Actually, a concept in both CCD and WordNet is just a SynSet (set of synonyms). The local structure (in a category) of CCD is a labeled tree and totally a network.*

*Philosopher L.Wittgenstein pointed in his book Philosophical Investigations, the similarities between any two concepts, whatever microcosmic or microscopical, form a network.

Principle of Substitution

Definition 2 Attributed to Leibniz, a SynSet $A = \{w_1, w_2, \cdots, w_n\}$ is well defined if there exists some context C such that $C(w_i/w_j)$* does not change the meaning (or truth value) of C (denoted by $[\![C(w_i/w_j)]\!] = [\![C]\!]$).

The identification of a SynSet has something to do with Corpus Linguistics, which will be discussed later.

*i.e., the substitution of w_i by w_j in context C, where w_i is in C.

Classification of Concepts

Definition 3 Essential concept was firstly proposed by Leibniz in his book Nouveaux Essais sur l'Entendment Humain to differentiate from those nonessential concepts like "bald" .* Essential concepts c and c' are able to be distinguished clearly despite of the experiences, while the non-essential concepts are not such case.† Essential concepts are, for instance, 人, 动物 and etc..

*The definition of "bald" once was a problem in philosophy — how many hairs does a man have to be a bandicoot? It is difficult to give an exact answer. That is, there is no clear boundary between the bald and non-bald.

†We have to describe the nonessential concepts by continuous fuzzy functions.

Symmetric and Transitive

Definition 4 If concept C is a binary relation, then C is

1. symmetric iff $xCy \to yCx$
2. transitive iff $xCy, yCz \to xCz$

Example 1 $C = \{$朋友$,$ 哥们儿$, \cdots\}$ is symmetric but not transitive. While the concept of 长辈 is transitive but not symmetric.

Induced and Initial Concepts

Definition 5 If there exists a function such that concept $C = f(C_1, C_2, \cdots, C_n)$, where $C_i, i = 1, 2, \cdots, n$ is a given concept, then C is called an induced concept of $C_1, C_2, \cdots, C_n$ in f way. A concept which is not induced is called initial.

Example 2 $C = \{$爷爷, 祖父$, \cdots\}$ is $\bar{C}^2$, where $\bar{C} = \{$父亲, 爹, 爸爸$\cdots\}$.

Example 3 Concept of 外祖父 is a $\bar{C}C'$, where C' is the concept of 母亲.

More tags

In the further work of CCD, essential concepts, induced concepts will be tagged. If necessary, the transitive and symmetric properties will be added in the deductive system of the concept network. CCD also includes other tags* related to Pragmatics inside a concept, such as, commendatory (derogatory), normal (abnormal), etc..

*See CCD Specification.

Labeled Tree

Definition 6 A labeled tree is a mathematical system $T = \langle\langle N, \preceq_H, \prec_P\rangle, Q, L\rangle$ satisfying:

1. N is a finite set of node
2. $\preceq_H$ is a partial ordering on N
3. $\prec_P$ is a strict partial ordering on N
4. $(\exists x \in N)(\forall y \in N)[x \preceq_H y], x$ is called root

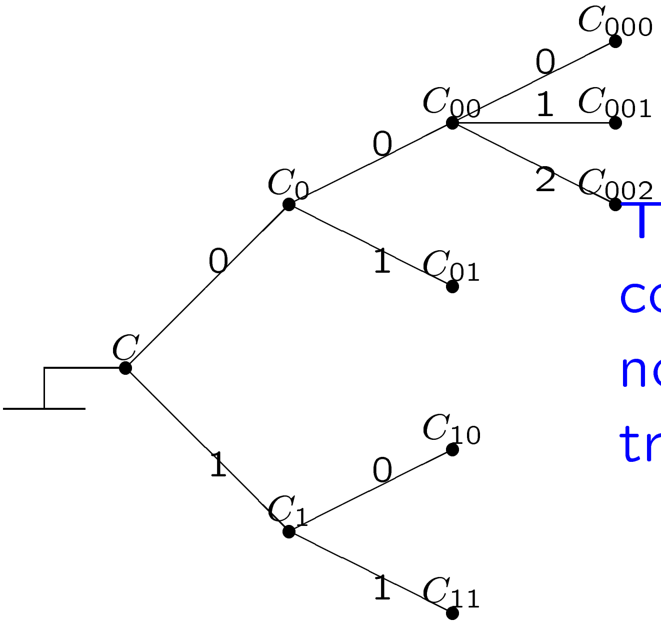
5. $(\forall x, y \in N)[[x \prec_P y \vee y \prec_P x] \leftrightarrow [x \not\preceq_H y \wedge y \not\preceq_H x)]$

6. $(\forall x, y, z, w \in N)[[w \prec_P x \wedge w \preceq_H y \wedge x \preceq_H z] \rightarrow y \prec_P z]$

7. Q is a finite set of labels

8. L : $N \rightarrow Q$ is a label map

Example of Labeled Tree



There is a method of coding for the set of nodes in the labeled tree.

Hypernymy Relation

Definition 7 If the proposition x is a kind of y is true, then y is the hypernym of x (denoted by $x \preceq_H y$) or x is the hyponym of y. Hypernymy relation is

1. transitive: $\forall x, y, z \in N, x \preceq_H y$ and $y \preceq_H z \to x \preceq_H z$

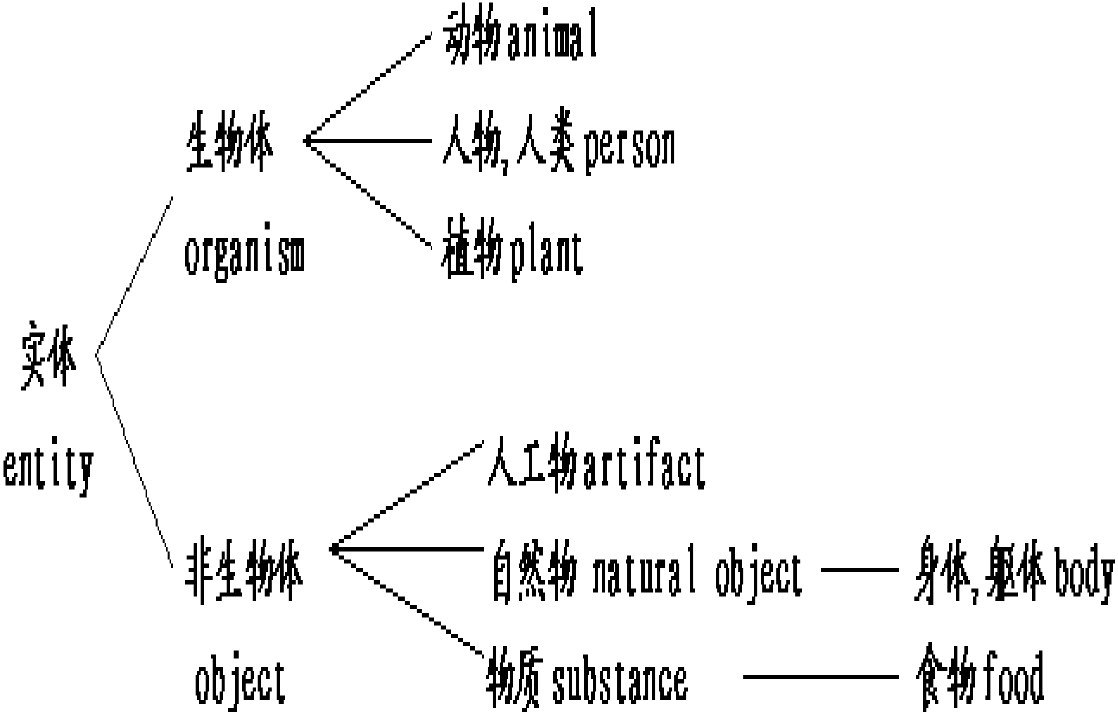
2. asymmetrical: $\forall x, y \in N, x \preceq_H y$ and $y \preceq_H x \to x = y$

Labeled Tree in CCD

In a specific category, hypernymy relation determines a hierarchical semantic structure.* CCD is a very special labeled tree because the label map in CCD is a bijection.

*Such hierarchical representations are widely used in the construction of IE (or IR) system, where they are called inheritance systems — a hyponym inherits all the features of the more generic concept and adds at least one feature that distinguishes it from its superordinate and from any other hyponyms of that superordinate. CCD is a concept inheritance system.

Hierarchical Structure



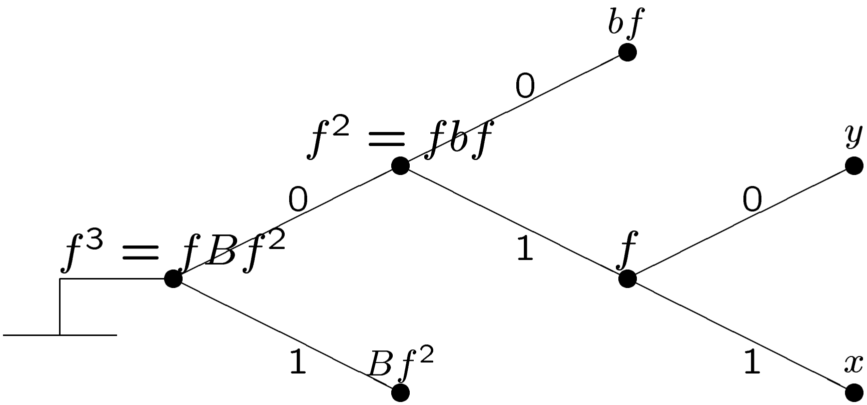
*

*Initial labeled tree given in both CCD and WordNet.

Operators in a Labeled Tree

Definition 8 Define three operators f, b and B denoting father, the nearest younger-brother and the nearest older-brother respectively. The nearest youngest-brother of y is the minimum of the set $\{x|y \prec_P x\}$, and its nearest oldest-brother is the maximum of set $\{x|x \prec_P y\}$. So, by the three operators, from a node x we can get to any higher node.

Example of Tree Operators



In the pedigree tree T_4, $011 = B(010)$ (or $010 = b(011)$), and $\forall x \in N, \exists f(x), b(x) \rightarrow f(x) = fb(x)$. Briefly, $f = fb$, also $f = fB$. Obviously, $f = fb \rightarrow f^2 = fbf, f = fB \rightarrow f^2 = fBf$ and $Bb = bB = 1$.

Burled Labeled Tree

Definition 9 Root is the unique node that has no father, and leaf is a node that has no son. If node x is not the root of T, and neither $B(x)$ nor $b(x)$ exists, then T is called burled.

Anyway, a burled labeled tree of concepts in a category is an unfinished inheritance system, although it is permitted during the generation of CCD.*

*The idea is from the discussion with Dr. Zhan Weidong.

Homomorphism and Isomorphism

Definition 10 $T_1 = \langle\langle N_1, \preceq_{H_1}, \prec_{P_1}\rangle, Q_1, L_1\rangle$ and $T_2 = \langle\langle N_2, \preceq_{H_2}, \prec_{P_2}\rangle, Q_2, L_2\rangle$ are two labeled trees, maps $N_1 \overset{\varphi}{\to} N_2$ and $Q_1 \overset{\psi}{\to} Q_2$ are defined by:

1. $x \preceq_{H_1} y \to \varphi(x) \preceq_{H_2} \varphi(y)$

2. $x \prec_{P_1} y \to \varphi(x) \prec_{P_2} \varphi(y)$

3. $\forall x \in N_1, \psi(L_1(x)) = L_2(\varphi(x))$

A tree-homomorphism from T_1 to T_2 is φ satisfying (1) and (2). A label-homomorphism from T_1 to T_2 is $\langle\varphi, \psi\rangle$ satisfying (1), (2) and (3). T_1 and T_2 are tree-isomorphic if there are tree-homomorphisms $T_1 \overset{\varphi}{\to} T_2$ and $T_2 \overset{\varphi'}{\to} T_1$ s.t. $\varphi'\varphi = \mathbf{1}_{N_1}$ and $\varphi\varphi' = \mathbf{1}_{N_2}$. A label-isomorphism $\langle\varphi, \psi\rangle$ between T_1 and T_2 is a label-homomorphism satisfying that φ is a tree-isomorphism and ψ is a bijection.

The intergradation of labeled tree T is label-homomorphic to the extension of T. Also, we can compare the structure of CCD with that of WordNet.

Brother Nodes in CCD

Definition 11 $x \in N$ is the oldest son if the offset of x is zero. $x \prec_P b(x) \prec_P \cdots \prec_P b^r(x)$ is the sequence of brother nodes. For instance, 春 $\prec_P$ 夏 $\prec_P$ 秋 $\prec_P$ 冬 or Sunday $\prec_P$ Monday $\prec_P \cdots \prec_P$ Saturday by time. The meaning of $\prec_P$ is evaluated by lexicographers.

Brother nodes are not ordered in WordNet, which means that WordNet is not a labeled tree because lacking of the restriction of strict partial ordering $\prec_P$ on N.

Distance in Labeled Tree

Definition 12 $\forall x, y \in N$, let $z \in N$ be their nearest ancestor satisfying $z = f^m(x)$ and $z = f^n(y), D(x,y) \overset{\text{def}}{=} m+n$. $k \in \mathbb{N}$ is called the offset of x from its oldest brother if $\exists B^k(x)$ and $\nexists B^{k+1}(x)$. Let the offset of y is l. Without generality, the degree of similarity between x and y is:

- If $mn = 1, S(x,y) \overset{\text{def}}{=} \langle 0, |k-l| \rangle$

- If $mn \neq 1, S(x,y) \overset{\text{def}}{=} \langle m+n, k+l \rangle$

Comparison of Similarities

Definition 13 Suppose that $S(x_1, y_1) = \langle a_1, b_1\rangle$ and $S(x_2, y_2) = \langle a_2, b_2\rangle$, the comparison of similarities is as follows:

1. $a_1 = a_2$
   - $S(x_1, y_1) \preceq S(x_2, y_2) \leftrightarrow b_1 \leq b_2$
2. $a_1 \neq a_2$
   - If $a_1 < a_2$, then $S(x_1, y_1) \prec S(x_2, y_2)$
   - If $a_1 > a_2$, then $S(x_2, y_2) \prec S(x_1, y_1)$

Theorem 1 $\langle\{S(x, y) | x, y \in N\}, \preceq\rangle$ is a totally ordered set.

Opposite Concepts

Definition 14 Two given concepts X and Y are opposite (denoted by $X = \neg Y$) iff $\exists C$ s.t. $[\![C(x/y)]\!] = [\![\neg C]\!]$, where $x \in X, y \in Y$ and $\frac{x}{X} \diamond \frac{y}{Y}$.*

Property 1 For any concept in CCD, there is at most one opposite concept, and $X = \neg Y \rightarrow Y = \neg X$. That is, the relationship between two opposite concepts is symmetric and unique.

* $\frac{x}{X} \diamond \frac{y}{Y}$ denotes that x and y are antonyms as the specific behaving of X and Y respectively.

Holonymy Relation

Definition 15 The part-whole relation between noun concepts (or verb concepts) is generally considered to be a semantic relation. Y is called a holonym[*] of X (comparable to opposite and hypernymy relations) if proposition X is a part of Y holds. Holonymy and hypernymy become intertwined in complex ways.

[*] For noun concepts, the holonymy relations in WordNet is a real subset of those in CCD (there are six distinct kinds of holonymy relations in CCD). See CCD Specification.

Temporal Classification

Definition 16 There are four distinct kinds of verb concepts:

temporal dependence	temporal independence
$V_1 \to \exists V_2[T(V_1) = T(V_2)]$	$V_1 \to \exists V_2[T(V_1) \prec T(V_2)]$
$V_1 \to \exists V_2[T(V_1) \subset T(V_2)]$	$V_1 \to \exists V_2[T(V_2) \prec T(V_1)]$

where $T(V)$ denotes the temporal interval (or point) of verb concept V. $V_1 \to \exists V_2[T(V_1) \subset T(V_2)]^*$ and $V_1 \to \exists V_2[T(V_2) \prec T(V_1)]$ (backward presupposition) are unified to entailment in WordNet, which are distinguished in CCD.

*Similar to holonymy relation between noun concepts.

Sentence Frame

The sentence frame for verb concepts in WordNet is very simple, which is improved to the type of longest chunk sequence in CCD. We are interested in the concept (or semantic) restrictions of a chunk and their statistical distribution.* Because the concept restrictions could be from other parts of CCD (for instance, noun concepts), the total structure of CCD is quite more complex than that of WordNet.

*This research is useful to the syntactic-semantic shallow parsing.

Dynamic Learning of CCD

Principle of Substitution, as a method to define a SynSet, is alway criticized to be much empirical. We have detected that Computational Lexicology could benefit from Corpus Linguistics and vice versa:

1. extension of tagged corpus and Hidden Markov Model (HMM) of concepts

2. machine learning of concept tags

3. rationality of SynSet and sentence frame

Extension of Tagged Corpus

Definition 17 Given a sentence template* $s = w_1 w_2 \cdots w_n$, each $w_i \in s$ corresponds to an unique SynSet $C_i = \{w_{i_1}, \cdots, w_{i_k}\}$. From the sequence of concepts $C_1 C_2 \cdots C_n$, the set of all possible sequences of words from the relevant SynSets is called the extension of sentence s.

The extension of corpus is defined similarly. One could find the size of the combinatorial extension is almost exponential.

*We just study nouns, verbs, adjectives and adverbs in a sentence.

HMM of Concepts

$s = \frac{w_1}{C_1}\frac{w_2}{C_2}\cdots\frac{w_n}{C_n}$ is a sentence tagged with concepts. Different from HMM of POSs,* the tags are well structured. To avoid the sparse data caused by too many tags, concept tags collapse along the labeled trees until HMM works (dynamic learning of concept tags), from which the chunk analysis will benefit. At the same time, HMM of concepts makes the automatic extension of corpus possible.

*See the draft Hidden Markov Model and its Applications in NLP in http://icl.pku.edu.cn/yujs/

Rationality of SynSet and Sentence Frame

To confirm a SynSet or a chunk sequence rational, it is necessary to check the extension of corpus tagged with concepts. The effects of semantic knowledge representation from the viewpoint of ontology, are dependent on the result of its applications.*

*In other words, there should be a dynamic learning of lexicon from the extension of corpus, which weakens the experiential factors in the definition of a SynSet.

Computational Lexicology and Corpus Linguistics

A great idea is to construct an Integrated Language Database (ILD) containing GKB, CSD, CCD and large corpus in ICL. Corpus is a very important resource in checking the details in the lexicons. We have an integrated automatic checking software to keep all the structures in CCD, and we will have a visualized auxiliary software to check the mistakes caused by knowledge representation.*

*With the increase of more concepts, the complexity of the lexicon will bring more terrible troubles, some are caused by the structure (such as cycle errors in WordNet), and others by knowledge representation.

Acknowledgement

Many thanks to Prof. Yu Shiwen for his trust in my understanding of CCD. And also I appreciate all the members participated in CCD project, especially Mr. Zhang Huarui, Ms. Song Chunyan, Mr. Liu Yang, Prof. Lu Chuan, Mr. Zhang Xuedong, Mr. Yu Yongbo and our advisor of linguistics, Dr. Zhan Weidong. The blithesome collaboration with Mr. Liu Dong and Ms. Wang Wei from Pecan is memorable for all of us. Lastly, thank Prof. Lu Chuan again for his kindly discussion in private with the author.

Thank you

for your attention!