Computational Linguistics and Chinese Language Processing
Vol. 14, No. 1, March 2009, pp. 1-18

Fertility-based Source-Language-biased Inversion Transduction Grammar for Word Alignment

Chung-Chi Huang* **and Jason S. Chang+**

Abstract

We propose a version of Inversion Transduction Grammar (ITG) model with IBM-style notation of fertility to improve word-alignment performance. In our approach, binary context-free grammar rules of the source language, accompanied by orientation preferences of the target language and fertilities of words, are leveraged to construct a syntax-based statistical translation model. Our model, inherently possessing the characteristics of ITG restrictions and allowing for many consecutive words aligned to one and vice-versa, outperforms the Bracketing Transduction Grammar (BTG) model and GIZA++, a state-of-the-art word aligner, not only in alignment error rate (23% and 14% error reduction) but also in consistent phrase error rate (13% and 9% error reduction). Better performance in these two evaluation metrics suggests that, based on our word alignment result, more accurate phrase pairs may be acquired, leading to better machine translation quality.

Keywords: Inversion Transduction Grammar, Syntax-based Statistical Translation Model, Word Alignment.

1. Introduction

A statistical translation model is a model which detects word correspondences within sentence pairs, whether relying on lexical information or on syntactic aspects of the involved languages or both. In spite of the fact that methodologies vary, the intention is clear: to obtain better word alignment results so that a better translation model implies better performance in different linguistic applications. Among the methodologies are phrase-based (Och & Ney, 2004; Chiang, 2005; Liu *et al*., 2006) and syntax-based machine translation systems (Galley *et al*., 2004; Galley *et al.*, 2006).

* CLCLP, TIGP, Academia Sinica, Taipei, Taiwan
+ Department of Computer Science, NTHU, Hsinchu, Taiwan
E-mail: {u901571; jason.jschang}@gmail.com

[Received June 6, 2008; Revised May 12, 2009; Accepted May 15, 2009]

Since the pioneering work of Brown *et al*. (1988), a myriad of research projects have focused on the statistical translation model. These could be classified into two main categories: one paying little attention to the grammar of the languages (Vogel *et al*., 1996; Och & Ney, 2000; Toutanova *et al*., 2002) and the other explicitly utilizing languages' structural or syntactic information (Wu, 1997; Yamada & Knight, 2001; Cherry & Lin, 2003; Gildea, 2004; Zhang & Gildea, 2005). With an increasing number of more accurate syntactic analyzers (*e.g*., part-of-speech tagger and Stanford parser) being developed and in view of the deficiency in modeling grammatical aspects of languages facing IBM-like models, the latter has received increasing attention.

Recently, in order to incorporate languages' syntax, Yamada and Knight (2001) transformed source-language (SL) (*e.g*., English) parse trees into target-language (TL) (*e.g*., Japanese) strings, using operations of reordering, inserting, and translating on tree nodes. Instead of accepting monolingual (*i.e*., SL or TL) parse trees to do the transformation, Wu's ITG model (1997) first associates production rules (*e.g*., S→NP VP) commonly shared by two languages with (straight or inverted) word orientation and, based on these synchronous rules, constructs bilingual parse trees at run time. This data-oriented parsing methodology is reported to outperform tree-to-string model (*i.e*., (Yamada & Knight, 2001)) concerning word-level alignment (Zhang & Gildea, 2004).

Even though the promising ITG is proposed, Wu (1997) conducts a word-aligning experiment leveraging a special case of ITG, minimal bracketing transduction grammar (BTG), in which languages' grammars are assumed to be unavailable, constituent categories (*e.g*., NP and VP) are not differentiated (using only three symbols: one for lexical translation rules, another for straight binary production rules, the other for inverted), and the probabilities of the straight and inverted binary rules are all assigned constant. These imply that the choices of straight or inverted word orientations would be made *solely* based on the bonds of lexical translations rather than on the structural divergences of the involved languages and that the potential of the syntax-oriented ITG would not be fully explored.

More recently, Zhang and Gildea (2005) presented a lexicalized BTG model where orientation choices are also dependent on the head words of the structural constituents. They expect lexical pairs passed up from the bottom (*i.e.*, leaf nodes) of the bilingual parse tree will make BTG models more knowledgeable in determining straight/inverted word order. Nonetheless, they found that lexical information at the lower levels of trees is more deterministic in word orientations than that at the higher levels.

To explore the power of ITG a little more (and inspired by Zhang *et al*. (2006), who suggest that binarized rules improve both speed and accuracy of a syntax-based machine translation system), in this paper, we describe a version of ITG model where the binary grammatical rules (*e.g*., S→NP VP) of the source language (*e.g*., English) are used as the

skeleton of our synchronous rules. Since the rules are biased toward the syntactic labels of the source language, our model is referred to as *bITG* model, short for biased ITG model. In our model, based on word-aligned sentence pairs, binary SL CFG rules are automatically annotated with the target language's word orientations and the associated orientation probabilities are automatically computed via Maximum Likelihood Estimation (MLE).

For example, take the languages of English, Chinese, and Japanese. The higher probability of our binary bITG rule VP→[VP NP], where the square brackets denote the same ordering (*straight*) of the two right-hand-side constituents in both languages when expanding the left-hand-side symbol, indicates a similar VO construct exists in English (SVO language) and Chinese (SVO language). On the contrary, the different VO construct in English and Japanese (SOV language) is modeled through the high *inverted* probability of the binary bITG rule VP→<VP NP> where the pointed brackets denote that we expand the left-hand-side symbol into two right-hand-side symbols in reverse orientation in two languages. Notice that these two bITG rules originate from *the same* binary CFG rule (VP→VP NP) of the source language, English, only with *different* ordering tendencies on the TL (*i.e.*, Chinese or Japanese) end.

In addition, we leverage IBM-style fertility probabilities of words to accommodate many-to-one or one-to-many word alignment links. In other words, in our model, many contiguous words in the source can be aligned to one word in the target and vice-versa. Originally, Wu's BTG model (1997) only allowed for a maximum of one-to-one word correspondences, which may affect the performance on word alignments and the accuracy of the bilingual parse trees. This one-to-one mapping restriction is especially not suitable for a language pair involving a language without clear word delimiters since the tokenization (or segmentation) of sentences of that language (*e.g.*, Chinese) prior to word alignment is independent of words of another (*e.g.*, English), resulting in tokens being under- or over-segmented for the corresponding words and, subsequently, abundant many-to-one/one-to-many word alignments.

The paper is organized as follows. Sections 2 and 3 describe our model in detail. Section 4 shows empirical results. Discussions are made before the conclusion in Section 6.

2. Method

In this section, we begin with an example of how bITG rules and fertilities of words are utilized to assist in word-aligning sentence pairs. Thereafter, a more formal description of our model will be discussed.

2.1 An Example

English sentence: These factors will continue to play a positive role after its return.

English POS tags: DT NNS MD VB TO VB DT JJ NN IN PRP$ NN

Chinese sentence: 香港 回歸 後 這些 條件 將會 繼續 發揮 積極 作用

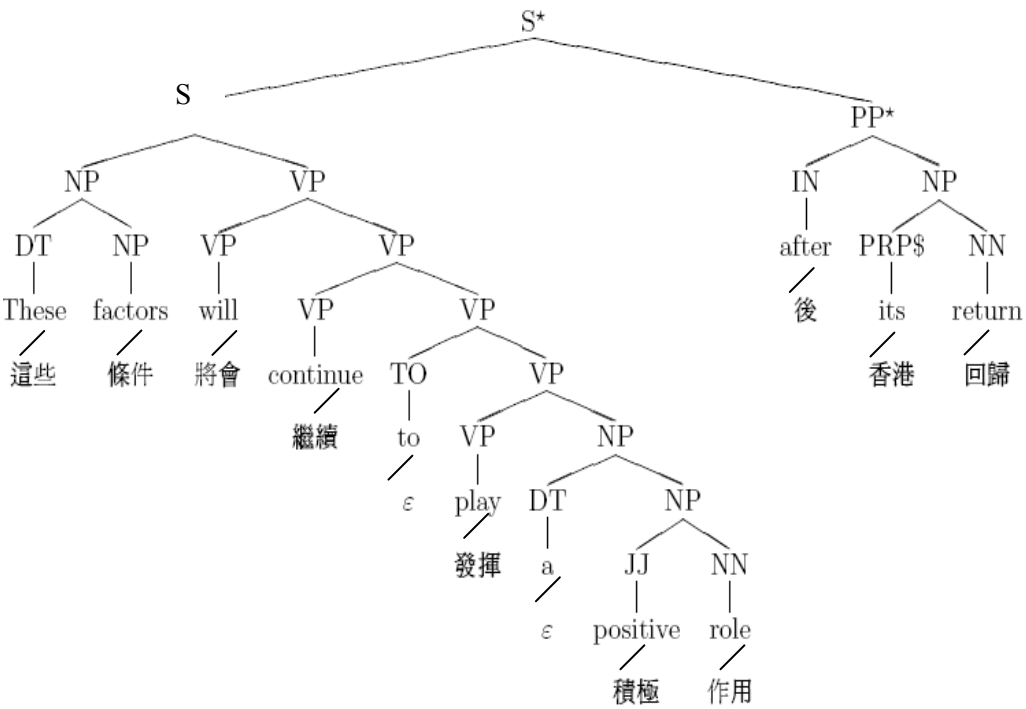

Figure 1. An example sentence pair and its bilingual parse tree

Once a sentence pair and the part-of-speech (POS) information of the SL sentence are fed into our model, it synchronously parses the sentence pair using unary lexical translation rules (*e.g.*, JJ→positive/積極 where / denotes word correspondence in two languages) and binary SL CFG rules attached with orientation preferences in the target language (*e.g.*, VP→[VP NP]). Also, the leaves of the bilingual parse tree are the word alignment results for this sentence pair.

During bilingual parsing, the model assigns probabilities to substring pairs of the bitext after each of them is associated with possible syntactic labels on the source side. For example, take the sentence pair and its parse in Figure 1, where spaces in the Chinese sentence are used to distinguish the boundaries of segments, ε stands for NULL, and * denotes the inverted orientation of the node's two children on the target. The substring pair (positive role, 積極 作用) associated with linguistic symbol NP will be assigned a probability. In this particular parse, the probability is the product of probabilities of the straight binary bITG rule, NP→[JJ NN],

and the lexical rules of bITG, JJ→positive/積極, and NN→role/作用. In our model, the higher probability of rule NP→[JJ NN] than the probability of the corresponding inverted rule NP→<JJ NN> does not merely instruct the model to align the two right-hand-side counterparts (*i.e.*, JJ and NN) of two languages in a straight fashion *more*, but also implies English and Chinese exhibit similar word-order regularity regarding the syntactic constituents.

On the other hand, in the example sentence pair, the beginning half, "*These factors will continue to play a positive role,*" is translated into the back of the Chinese sentence whereas the ending half, "*after its return,*" is translated into the beginning. Inverted rules (*e.g.*, S→<S PP>) are designed to capture such systematic differences in the languages' grammars.

What is more, since only monolingual information is exploited to segment Chinese sentences, it is likely that the word alignments will *not* be constrained to one-to-one, one-to-zero, and zero-to-one mappings. For instance, *香港* is often segmented as *a* word in Chinese but needs to be aligned to two words (*Hong* and *Kong*) in English, a case of two-to-one mapping. Therefore, we incorporate notion of fertility into our model.

As for the example of "*Hong Kong*" aligned to "*香港*", three possible word-aligning scenarios concerning fertility will be considered at runtime parsing: zero fertility of *Hong* and singular fertilities of *Kong* and *香港* where *Hong* is aligned to NULL but *Kong* is aligned to *香港*; zero fertility of *Kong* and singular fertilities of *Hong* and *香港* where *Kong* is aligned to NULL but *Hong* is aligned to *香港*; singular fertilities of *Hong* and *Kong* and dual fertility of *香港* where both *Hong* and *Kong* are aligned to *香港*.

Taking into account the probabilities of lexical translations, binary grammatical rewrite rules, and fertilities of words, our model manages to find a better parse tree that applies more appropriate synchronous rules to match the structural divergences and more suitable lexical mapping relations (one-to-one, one-to-two, *et al.*) in two languages. Better parses are more likely to yield better word alignment results.

We actually estimate the probabilities of bITG rules, consisting of unary lexical translation rules and binary SL CFG rules with word orientation on the TL, and those of the fertilities of words from a parallel corpus and an SL CFG. We will discuss the training algorithm in more detail in Section 3.

2.2 Formal Description

We now formally describe our statistical translation model. To be comparable to previous work, the English-French notation is used throughout this paper. *E* and *F* denote the source and target language, respectively, and e_i stands for the *i*-th word in sentence *e* in language *E* and f_j for the *j*-th word in sentence *f* in *F*.

Given $(e, f) = (e_1 \cdots e_m, f_1 \cdots f_n)$ and the POS tag sequence of e, τ, our model aims

to construct the most probable bilingual parse tree B_t^*, satisfying $\arg\max_{B_t}\{\Pr(B_t|e, f, \tau)\}$, with the by-product of word-level correspondences. Intuitively, the probability of a bilingual parse tree B_t provided with *e*, *f*, and τ is modeled as the product of probabilities associated with grammatical rewrite rules and lexical information:

$$\Pr(B_t|e, f, \tau) = \Pr(\mathbf{D}|e, f, \tau) \times \Pr(\mathbf{A}|e, f, \tau) \quad (1)$$

where, by inspecting the parse tree B_t, **D,** and **A** represent the set of its production rules with syntactic labels on the right hand side (*e.g.*, NP→JJ NN) and the set of rules with word alignments on the right (*e.g.*, JJ→positive/積極), respectively.

For simplicity, we use α_k to denote internal nodes (NP, JJ, *etc*) of the tree B_t, whereas we use β_k to denote leaf nodes (*e.g.*, these/這些, positive/積極). Tree nodes in B_t can be divided into three groups according to the number of children they are connected to: the first, denoted by set $\mathbf{N}_2$, consists of nodes with two children; the second, denoted by set $\mathbf{N}_1$, is made up of nodes with one child; the last, denoted by set $\mathbf{N}_0$, comprises nodes without a child. For notation convenience, each $\alpha_k \in \mathbf{N}_2$ has two children represented by α_{2k} and α_{2k+1}, and each $\alpha_k \in \mathbf{N}_1$ has one child β_k.

In our model, the probability of constructing B_t is the product of the probabilities of two sources: the first estimating the probabilities of the applied binary bITG rules; the second estimating those of the unary lexical translation rules and the fertilities of words in the tree. Assuming each applied rule is independent of one another, we rewrite the grammatical-related term in Equation (1) as

$$\Pr(\mathbf{D}|e, f, \tau) \cong \prod_{\alpha_k \in \mathbf{N}_2} \mathrm{P}^{\lambda_1}(\alpha_k \rightarrow [\![\alpha_{2k}\ \alpha_{2k+1}]\!]) \quad (2)$$

where $[\![\ ]\!]$ can be straight $[\ ]$ or inverted $\langle\ \rangle$. On the other hand, the lexical-related term in Equation (1) is decomposed into three factors, as shown in Equation (3): one for the product of probabilities of lexical translation rules given τ, another for the product of fertility probabilities of words in *e,* and the other for the product of fertility probabilities of words in *f*.

$$\Pr(\mathbf{A}|e, f, \tau) \cong \prod_{\alpha_k \in \mathbf{N}_1} \mathrm{P}^{\lambda_2}(\alpha_k \rightarrow \beta_k|\tau) \times \prod_{i=1}^{m} \mathrm{P}^{\lambda_2}(\Phi = \phi_{e_i}) \times \prod_{j=1}^{n} \mathrm{P}^{\lambda_2}(\Phi = \phi_{f_j}) \quad (3)$$

In Equation (3), Φ is the random variable for fertilities of words, and ϕ_{e_i} and ϕ_{f_j} denote fertilities of e_i and f_j, respectively. From Equations (1) to (3), we estimate the probability of a parse tree via

$$\Pr(B_t|e,f,\tau) \cong \prod_{\alpha_k \in \mathbf{N_2}} \mathrm{P}^{\lambda_1}\left(\alpha_k \to [\![\alpha_{2k}\ \alpha_{2k+1}]\!]\right) \times \prod_{\alpha_k \in \mathbf{N}_1} \mathrm{P}^{\lambda_2}\left(\alpha_k \to \beta_k|\tau\right) \times \prod_{i=1}^{m} \mathrm{P}^{\lambda_2}\left(\Phi_{e_i} = \phi_{e_i}\right) \times \prod_{j=1}^{n} \mathrm{P}^{\lambda_2}\left(\Phi_{f_j} = \phi_{f_j}\right) \quad (4)$$

in which the sum of the weight λ_1 and λ_2 is one.

2.3 Runtime Parsing

In this subsection, we depict a CYK-like parsing algorithm for obtaining the most likely bilingual parse tree given the sentence pair $(e,f)=(e_1\cdots e_m, f_1\cdots f_n)$, the pre-determined POS tag sequence, $(t_1,\cdots,t_m)$, of sentence *e*, and the grammar *G* in *E* (*i.e.*, SL grammar). Notice that our model is a data-driven one as is Wu (1997). In other words, it synchronously parses sentence pair via bITG rules *without* a monolingual (SL or TL) parse tree. Figure 2 shows the run-time parsing algorithm.

Parsing Algorithm

//Initial Step

For $1 \le i \le m, 1 \le j \le n$

(1) $\delta_{t_i,i-1,i,j-1,j} = \mathrm{P}^{\lambda_2}\left(t_i \to e_i/f_j\right) \times \mathrm{P}^{\lambda_2}\left(\Phi_{e_i}=1\right) \times \mathrm{P}^{\lambda_2}\left(\Phi_{f_j}=1\right)$

(2) For every $L \to t_i \in G$ in E

(3) $\delta_{L,i-1,i,j-1,j} = \mathrm{P}^{\lambda_2}\left(L \to e_i/f_j\right) \times \mathrm{P}^{\lambda_2}\left(\Phi_{e_i}=1\right) \times \mathrm{P}^{\lambda_2}\left(\Phi_{f_j}=1\right)$

For $1 \le i \le m, 0 \le j \le n$

(4) $\delta_{t_i,i-1,i,j,j} = \mathrm{P}^{\lambda_2}\left(t_i \to e_i/\varepsilon\right) \times \mathrm{P}^{\lambda_2}\left(\Phi_{e_i}=0\right)$

(5) For every $L \to t_i \in G$ in E

(6) $\delta_{L,i-1,i,j,j} = \mathrm{P}^{\lambda_2}\left(L \to e_i/\varepsilon\right) \times \mathrm{P}^{\lambda_2}\left(\Phi_{e_i}=0\right)$

For $0 \le i \le m, 1 \le j \le n, L \in$ syntactic labels on E end

(7) $\delta_{L,i,i,j-1,j} = \mathrm{P}^{\lambda_2}\left(L \to \varepsilon/f_j\right) \times \mathrm{P}^{\lambda_2}\left(\Phi_{f_j}=0\right)$

//Recurrent Step

For any possible (s,t,u,v) // $1 \le s,t \le m, 1 \le u,v \le n$

For any possible grammatical label *p*

If ($t \ge s$ and $v \ge u$) and not ($t = s$ and $v = u$)

(8)
$$\delta_{p,s,t,u,v} = \max_{\substack{q,r\in\text{syntax labels on } E \\ s\le s'\le t \\ u\le u'\le v}} \left\{ \begin{array}{l} \mathrm{P}^{\lambda_1}\left(p \to [q\ r]\right) \times \delta_{q,s,s',u,u'} \times \delta_{r,s',t,u',v}, \\ \mathrm{P}^{\lambda_1}\left(p \to \langle q\ r \rangle\right) \times \delta_{q,s,s',u',v} \times \delta_{r,s',t,u,u'} \end{array} \right\}$$

//for backtracking

$$\mathrm{b}_{p,s,t,u,v} = \arg\max_{\substack{q,r\in\text{syntax labels on } E \\ s\le s'\le t \\ u\le u'\le v}} \left\{ \begin{array}{l} \mathrm{P}^{\lambda_1}\left(p \to [q\ r]\right) \times \delta_{q,s,s',u,u'} \times \delta_{r,s',t,u',v}, \\ \mathrm{P}^{\lambda_1}\left(p \to \langle q\ r \rangle\right) \times \delta_{q,s,s',u',v} \times \delta_{r,s',t,u,u'} \end{array} \right\}$$

(9) Backtrack()

Figure 2. Run-time parsing.

During a parse of a sentence pair in our model, a table of $\delta_{p,s,t,u,v}$, the *best* probability for parsing substring pair $(e_{s+1}\cdots e_t, f_{u+1}\cdots f_v)$ attached with a syntactic symbol p on E side, is constructed.

In Step (1) of Figure 2, we compute the probability of a one-to-one word correspondence e_i/f_j with e_i's pre-determined POS tag t_i, according to the probability of the unary bITG rule $t_i \to e_i/f_j$ and the probabilities of fertilities of e_i and f_j (fertilities are 1s for one-to-one mapping). Since the POS tag t_i can be derived from some possible phrasal constituents in *G* (Step (2)) (*e.g.*, NN can be derived from NP), we also compute their associated probabilities (Step (3)). Similarly, in Steps (4) to (7), we calculate the probabilities of the one-to-zero and zero-to-one word correspondences limited to the scope of the sentence pair.

Afterwards, relying on the work done previously, word correspondences and parsing results of longer substring pairs would unveil themselves in a bottom-up manner. In Step (8), *s*' divides the substring $e_{s+1}\cdots e_t$, labeled as *p*, into two parts, $e_{s+1}\cdots e_{s'}$ and $e_{s'+1}\cdots e_t$, with *q* as a possible grammatical symbol of the first part and *r* as a possible symbol of the second, while *u*' divides the substring $f_{u+1}\cdots f_v$ into $f_{u+1}\cdots f_{u'}$ and $f_{u'+1}\cdots f_v$. As the substring $e_{s+1}\cdots e_{s'}$ can be aligned to $f_{u+1}\cdots f_{u'}$ or $f_{u'+1}\cdots f_v$, both straight and inverted orientation of the SL CFG rules "$p \to q\ r$" ought to be considered. Note that the computation in Step (8) does not properly deal with the cases of many-to-one or one-to-many word-level alignments. For many-to-one alignments, $\delta_{p,s,t,u-1,u}$ should further incorporate the parsing candidate:

$$\mathrm{P}^{\lambda_2}\left(\Phi_{f_u} = (t-s)\right) \times \max_{\substack{q,r\in\text{syntax} \\ \text{labels on } E}} \left\{ \mathrm{P}^{\lambda_1}\left(p \to [q\ r]\right) \times \frac{\delta_{q,s,s+1,u-1,u}}{\mathrm{P}^{\lambda_2}\left(\Phi_{f_u} = 1\right)} \times \frac{\delta_{r,s+1,t,u-1,u}}{\mathrm{P}^{\lambda_2}\left(\Phi_{f_u} = (t-s-1)\right)} \right\}$$

where $\delta_{r,s+1,t,u-1,u}$ needs to be constructed from many-to-one or one-to-one word mapping relation since words $e_{s+1}\cdots e_t$ are all aligned to f_u. A similar principal applies to one-to-many mapping (*i.e.*, the calculation of $\delta_{p,s-1,s,u,v}$).

Finally, using the standard CYK backtracking technique, we can find the most probable bilingual parse tree of the sentence pair with word alignment results. The integration of fertilities of words into the model aims to improve the parsing and the word-aligning quality.

2.4 Pruning

Although the complexity of the described algorithm is polynomial-time (proportion to m^3n^3), the execution time grows rapidly with the increase in the variety of syntactic labels, from three structural labels (Wu, 1997) to the grammatical categories of the source language's syntax in our model. As a result, pruning techniques are essential to reduce the time spent on parsing.

We adopt pruning in the following two manners. The first pruning technique is, for a given SL substring $e_{s+1}\cdots e_t$ and a given TL substring's length, to only keep parse trees whose probabilities fall within the *best* $N\times\sigma$, where N is the number of possible parses for a SL substring $e_{s+1}\cdots e_t$ and a length of the TL substring, and σ is a real number between 0 and 1. In other words, we remove inferior parse trees that are not in the set of the best $N\times\sigma$ ones. Since *N* varies from case to case (depending on the SL substring and the length of TL substring), only the more probable trees within the ratio (*i.e.*, σ) of *N* will remain.

The second pruning technique is related to the ratio of the length of the SL and TL substring. $\delta_{p,s,t,u,v}$ will not be calculated if $\frac{t-s}{v-u}$ is smaller than θ_{ratio} or larger than $1/\theta_{ratio}$ where $0\le\theta_{ratio}\le 1$, since few words will be aligned to more than $1/\theta_{ratio}$ words in another language.

By applying the aforementioned pruning techniques, the time spent on parsing each sentence pair can be reduced by *more than* half. Empirically, pruning unlikely parses has little affect on the word alignment quality but reduces computational overhead significantly.

3. Probability Estimation

In this section, we describe how to estimate the probabilities of our unary bITG rules (*e.g.*, JJ→positive/積極) and binary bITG rules (*e.g.*, VP→[VP NP]) which denote the association of bilingual lexical words and model the structural divergences of the two languages, respectively. Figure 3 shows the probabilistic estimation procedure.

(1) $\mathbf{H} = \mathbf{WA}$

(2) For $\left(r, e_{i_1}^{i_2}, f_{j_1}^{j_2}, L, rhs, rel\right) \in \mathbf{H}, \left(r, e_{i_1'}^{i_2'}, f_{j_1'}^{j_2'}, L', rhs', rel'\right) \in \mathbf{H}$ have not been considered

(3) If $\left(i_2 = i_1' - 1\right)$

(4) For every $L'' \rightarrow L\ L' \in G$ in E

(5) If $\left(j_2 + 1 \le j_1' \le j_2 + \delta\right)$

(6) $\mathbf{H} = \mathbf{H} \cup \left\{\left(r, e_{i_1}^{i_2'}, f_{j_1}^{j_2'}, L'', L\ L', Straight\right)\right\}$

(7) If $\left(j_2' + 1 \le j_1 \le j_2' + \delta\right)$

(8) $\mathbf{H} = \mathbf{H} \cup \left\{\left(r, e_{i_1}^{i_2'}, f_{j_1'}^{j_2}, L'', L\ L', Inverted\right)\right\}$

(9) similar strategy applies when $i_2' = i_1 - 1$

Incorporate words aligned to null, each of which is denoted using 6-tuple representation, in both languages into **H**.

For $\left(r, e_{i_1}^{i_2}, f_{j_1}^{j_2}, L, rhs, rel\right) \in \mathbf{H}$

If (rhs is not a lexical pair)

(10) $$\mathrm{P}\left(L \rightarrow \left[R_1\ R_2\right]\right) = \frac{\mathrm{count}\left(\left(*,*,*,L,R_1\ R_2, Straight\right); \mathbf{H}\right)}{\left|\mathbf{H}\right|}$$

(11) $$\mathrm{P}\left(L \rightarrow \left\langle R_1\ R_2\right\rangle\right) = \frac{\mathrm{count}\left(\left(*,*,*,L,R_1\ R_2, Inverted\right); \mathbf{H}\right)}{\left|\mathbf{H}\right|}$$

Else

(12) $$\mathrm{P}\left(L \rightarrow e_i / f_j\right) = \frac{\mathrm{count}\left(\left(*,*,*,L,e_i / f_j, *\right); \mathbf{H}\right)}{\left|\mathbf{H}\right|}$$

(13) Based on **WA** and **C**, estimate $\mathrm{P}\left(\Phi_{e_i}\right)$ and $\mathrm{P}\left(\Phi_{f_j}\right)$ via MLE

Figure 3. The procedure of probabilistic estimation.

In Step (1) of our training procedure, an existing word-aligning strategy or tool (*e.g.*, GIZA++) is employed to obtain the word alignments (*i.e.*, **WA**) of a parallel corpus **C**. **WA** comprises elements of the form $\left(r, e_{i_1}^{i_2}, f_{j_1}^{j_2}, L, rhs, rel\right)$, which represents that the substring pair $\left(e_{i_1} \cdots e_{i_2}, f_{j_1} \cdots f_{j_2}\right)$ in sentence pair r has $L \rightarrow rhs$ as the production rule leading to the bilingual structure and has *rel* (either *straight* or *inverted*) as the cross-language word-order relation of the constituents of *rhs*. *rhs* denotes either a sequence of syntactic labels or a terminating bilingual word pair. Following this format, the example parses of (positive,積極)$_{JJ}$ and (after its return,香港 回歸 後)$_{PP}$ in Figure 1 would be denoted by the 6-tuple (193, e_8^8, f_9^9 ,JJ,positive/積極,don't_care) and (193, e_{10}^{12}, f_1^3 ,PP,IN NP,*Inverted*) respectively, where 193 is the record number of this sentence pair.

Then, we recursively select two sections of a sentence pair, which have not yet been

paired up, from **H** (Step (2))**.** If the SL substring of the first section (*i.e.*, $e_{i_1}^{i_2}$) is adjacent to that of the second (*i.e.*, $e_{i_1'}^{i_2'}$) on the right (Step (3)), based on word alignment result (Step (5) and Step(7)), a new straight-ordered (Step (6)) or inverted-ordered (Step (8)) section representing these two will be added into **H**. Specifically, once the SL substrings are related to some possible binary SL CFG rules, the right-hand-side constituents of these rules will be associated with an orientation on the TL end based on word alignment links. Since our model is a synchronous bilingual parsing one, *without* a monolingual parse tree, it enumerates all possible syntactic symbols to derive L and L' in Step (4). Note that, in Steps (5) and (7), δ, a small positive integer, is utilized to tolerate aligning errors introduced by the automatic word aligner or explicitness issue[1] during translation from one language to another, when determining cross-language straight/inverted word order phenomenon.

From Step (10) to Step (12), in which |**W**| stands for the number of entries in set **W** and count(*p*;**Q**) for the frequency of *p* in set **Q**, we estimate probabilities of bITG rules via Maximum Likelihood Estimation. In our model, the probabilities of lexical translation rules (*e.g.*, JJ→positive/積極) and binary bITG rules (*e.g.*, VP→[VP NP]) are estimated from the same source (*i.e.*, **H**). Alternative probabilistic estimation of these two kinds of rules can be adopted. For example, the probabilities of lexical translation rules can be derived from pure word alignment set **WA** while those of binary bITG rules can be derived from set **H** without word-level alignment links. We employ the former estimation approach and, in experiments, it yields satisfying results (see Section 4), suggesting word-order tendencies of the two languages are properly modeled.

Finally, fertility probabilities related to words in both languages are also calculated (Step (13)).

4. Experiments

In experiments, we trained our model on a large English-Chinese parallel corpus. We examined word alignments produced by our bITG model using the evaluation metrics proposed by Och and Ney (2000). For comparison, we also trained GIZA++, a state-of-the-art word-aligning system, on the same parallel corpus.

4.1 Training Proposed Model

We used the news portion of Hong Kong Parallel Text[2] (HKPT) distributed by Linguistic Data Consortium as our sentence-aligned corpus **C,** which consisted of 739,919 English-Chinese

[1] Some translations may be omitted for conciseness, or some of the function words in one language may have no counterparts in another.

[2] LDC2004T08

sentence pairs. The average length was 24.4 words for English and 21.5 words for Chinese.

In our model, English sentences were considered to be the source while Chinese sentences were the target. SL sentences were POS tagged and TL sentences were segmented prior to word alignment. During training (as described in Section 3), we employed a GIZA++ run with default settings to obtain the word alignment set **WA** and our binary SL CFG *G* was based upon PTB section 23[3] production rules distributed by Andrew B. Clegg.

4.2 Evaluation

To evaluate our statistical translation model, 114 sentence pairs were chosen randomly from the news portion of HKPT as our testing data set. For the sake of execution time, we only selected sentence pairs whose SL and TL length did not exceed 15. Sentence pairs satisfying such a length constraint covered approximately 40% of the sentence pairs in the news portion of HKPT and were expected to be better word aligned via GIZA++.

We examined the word-aligning performance using the metrics of alignment error rate (AER) proposed by Och and Ney (2000), in which the quality of a word alignment result **A** produced by an automatic system is evaluated by:

$$precision = \frac{|\mathbf{A} \cap \mathbf{P}|}{|\mathbf{A}|}, \quad recall = \frac{|\mathbf{A} \cap \mathbf{S}|}{|\mathbf{S}|} \text{ and } \quad AER(\mathbf{S}, \mathbf{P}; \mathbf{A}) = 1 - \frac{|\mathbf{A} \cap \mathbf{S}| + |\mathbf{A} \cap \mathbf{P}|}{|\mathbf{A}| + |\mathbf{S}|}.$$

In AER, **S** (sure) denotes the set whose alignments are not ambiguous and **P** (possible) denotes the set consisting of alignments that might or might not exist $(\mathbf{S} \subseteq \mathbf{P})$. Thus, human annotations may contain many-to-one, one-to-many, or even many-to-many word alignments. Table 1 shows the experimental results of GIZA++, the BTG model (Wu, 1997), and our fertility-based SL-biased ITG model.

Table 1. Results of test data of different systems

	P	R	AER	F
E to F	**.891**	.385	.459	.537
F to E	.882	.533	.333	.664
Refined	.879	.635	.261	.737
BTG	.844	.610	.290	.708
bITG w/o fertility	.866	.638	.263	.735
bITG w/ fertility	.878	**.692**	**.224**	**.774**

[3] http://textmining.cryst.bbk.ac.uk/acl05/

In this table[4], P, R, and F stand for precision, recall, and F-measure[5], respectively. The performance of the E-to-F alignments (E stands for English and F for Chinese), the F-to-E alignments, and the refined alignments (proposed by Och and Ney (2000)) from both E-to-F and F-to-E directions of GIZA++ are shown in first three rows, along with that of BTG, which also trained on the word-aligning output of GIZA++. The results of our translation model *without* or *with* the capability of making many-to-one/one-to-many links are listed in the last two rows.

Compared with the BTG model that *does not* distinguish the constituent categories and makes the orientation choices merely on lexical evidence (without the information of languages' grammars), our model *without* fertility probability which allows for at most one-to-one alignment, as the BTG model does, achieved 9% reduction in the alignment error rate. This indicates that the binary SL CFG rules encoding with TL ordering preference in our model do capture the linguistic information of the languages such as word-order regularities or grammar and do impose more realistic and accurate reordering constraints on word alignment in the language pairs.

Furthermore, in comparison to the refined alignments of both word-aligning directions, our model *with* the concept of fertility (allowing for many-to-one/one-to-many links), which is quite similar to the refined approach accommodating many-to-many word mappings, increased the recall by 9% while maintaining high precision and achieved 14% alignment error reduction overall (increased F-measure by 5%).

As suggested by Table 1, it is safe to say that the proposed model yields more accurate bilingual parse trees, thus better word alignment quality, by introducing binary CFG rules of a language (*i.e.*, the source language) and fertility notation of IBM models into ITG model.

5. Discussion

In this section, we examine how the learnt similarities (*straight*) and differences (*inverted*) in word orders of two languages aid the word-aligning process of our model by means of the adjacency feature and cohesion constraint, mentioned in Cherry and Lin (2003). Subsequently, to evaluate the possibility of better machine translation quality by providing our model's output (*i.e.*, word correspondences), we adopt the recently-proposed metric, consistent phrase error rate (CPER) by Ayan and Dorr (2006).

---

[4] $\frac{|\mathbf{S}|}{|\mathbf{P}|}$ is 85.56% in human-annotated test data.

[5] Calculated using the formula $2 \times \mathrm{P} \times \mathrm{R} / (\mathrm{P}+\mathrm{R})$.

5.1 *Straight/Inverted* Orientation

Table 2 shows the accuracy of adjacent alignments made by our model, and the accuracy achieved by the refined approach is shown for comparison. If compared against the gold standard in the sure set (*i.e.*, **S** in Section 4), our model with bITG rules relatively increased the accuracy by more than 3%, suggesting the similar (or straight) word orientations of the binary syntactic constituents (*e.g.*, JJ and NN) in the languages are better captured in our model than in GIZA++. Note that alignments must have orders before an adjacency feature exists (see Cherry and Lin (2003)) in them. Therefore, an ordering, depending on the position of the English word in the sentence, was imposed to examine the feature.

Table 2. Examination of adjacent links

	Compared tosure links	Compared to possible links
Refined	.835	.869
bITG w/ fertility	**.863**	**.881**

Additionally, we examined whether the inverted binary bITG rules captured the diversities of the two grammars and helped to make correct crossing (or reverse) alignment links or not. For that purpose, we first acquired the dependency relations of the source (*i.e.*, English) sentences via a Stanford parser, and computed the percentage of links violating the cohesion constraint (see Cherry and Lin (2003)). The ratios of having crossing dependencies in the mapped Chinese dependency trees[6] are summarized in Table 3. As suggested by Table 3, our model reduced sixteen percent of the links violating the cohesion constraint (compared to the refined approach).

Table 3. Percentage of links violating cohesion constraint

	Percentage
Refined	.044
bITG w/ fertility	**.037**

The above statistics indicate that the probabilities related to straight and inverted word orders of bITG rules in our model not only impose a more suitable alignment constraint but properly model the systematic similarities and differences in two languages' grammars.

[6] Chinese dependency trees are mapped from English dependency trees based on word correspondences.

5.2 CPER

According to Ayan and Dorr (2006), the intrinsic evaluation metric of AER (Och and Ney, 2000) examines only the quality of word-level alignments and correlates poorly with the MT-community metric—BLEU score. As a result, we exploited consistent phrase error rate (CPER) to evaluate words alignments in the context of machine translation. CPER is reported to better correlate with translation quality (the smaller the CPER is, the better the translation quality) in that it evaluates phrase-level alignments and in that phrase-level alignments (bilingual phrase pairs) constitute the key essences of a MT system.

In Ayan and Dorr (2006), precision (P), recall (R), and CPER are computed via:

$\mathrm{P} = \frac{|P_A \cap P_G|}{|P_A|}$, $\mathrm{R} = \frac{|P_A \cap P_G|}{|P_G|}$, and $\mathrm{CPER} = 1 - \frac{2 \times \mathrm{P} \times \mathrm{R}}{\mathrm{P} + \mathrm{R}}$ where P_A and P_G stand for two sets of phrases generated by an automatic alignment *A* and manual alignment *G*, respectively. In Table 4, the proposed fertility-based source-language-based ITG model yielded the lowest CPER. This indicates that MT systems, accepting our word alignment output, are more likely to lead to better translation performance.

Table 4. Reports on CPER

	P	R	CPER
E to F	.479	.383	.574
F to E	.544	.518	.470
Refined	.573	.606	.411
BTG	.569	.569	.431
bITG w/o fertility	.598	.597	.402
bITG w/ fertility	**.624**	**.626**	**.375**

6. Conclusion and Future Work

To combine the strengths of the competing models, a thought-provoking fusion of IBM-style fertility with syntax-based ITG model is described. In our model, the orientation probabilities of the binary SL-based ITG rules are automatically estimated based on a word-aligned parallel corpus and are devised to better capture structural divergences of the involved languages. The proposed bITG model with fertility reduces AER by 14% and 23%, and reduces CPER by 9% and 13% compared to GIZA++ and Wu's BTG (1997), respectively. Lower CPER suggests MT systems chained after our statistical translation model are likely to yield better translation quality. In this paper, the performance of ITG models trained on large-scale bitexts is shown

for the first time with quite encouraging results.

As for future work, we would like to explore methods (*e.g*. (Brown, 1992)) for partitioning long sentences into shorter ones so that the time spent on bilingual parsing in our model can be reduced. We also like to see whether word-aligning quality can be further improved if our bITG rules are lexicalized, especially when lexical contents play an important role in determining word orders of the languages.

References

Ayan, N. F. & Dorr, B. J. (2006). Going beyond AER: an extensive analysis of word alignments and their impact on MT. In *Proceedings of ACL-2006*, 9-16.

Brown, P. F., Pietra, S. A. D., Pietra, V. J. D., Mercer, R. L., & Mohanty, S. (1992). Dividing and conquering long sentence in a translation system. In *Proceedings of the Workshop on Speech and Natural Language*, 267-271.

Brown, P. F., Pietra, S. A. D., Pietra, V. J. D., & Mercer, R. L. (1993). The mathematics of statistical machine translation: parameter estimation. *Computational Linguistics*, 19(2), 263-311.

Cherry, C. & Lin, D. (2003). A probability model to improve word alignment. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, 88-95.

Chiang, D. (2005). A hierarchical phrase-based model for statistical machine translation. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics*, 263-270.

Clegg, A. B. & Shepherd, A. (2005). Evaluating and integrating Treebank parsers on a biomedical corpus. In *Association for Computational Linguistics Workshop on software 2005*.

Galley, M., Hopkins, M., Knight, K., & Marcu, D. (2004). What's in a translation rule? In *Proceedings of HLT/NAACL-2004*, 273-280.

Galley, M., Graehl, J., Knight, K., Marcu, D., DeNeefe, S., Wang, W. et al. (2006). Scable inference and training of context-rich syntactic translation models. In *Proceedings of the 44th Annual Conference of the Association for Computational Linguistics*, 961-968.

Gildea, D. (2004). Dependencies vs. constituents for tree-based alignment. In *Proceedings of the EMNLP*, 214-221.

Liu, Y., Liu, Q., & Lin, S. (2006). Tree-to-string alignment template for statistical machine translation. In *Proceedings of the 44th Annual Conference of the Association for Computational Linguistics*, 609-616.

Och, F. J. & Ney, H. (2000). Improved statistical alignment models. In *Proceedings of the 38th Annual Conference of ACL-2000*, 440-447.

Och, F. J. & Ney, H. (2004). The alignment template approach to statistical machine translation. *Computational Linguistics*, 30(4), 417-449.

Toutanova, K., Ilhan, H. T., & Manning, C. D. (2002). Extentions to HMM-based statistical word alignment models. In P*roceedings of the Conference on Empirical Methods in Natural Processing Language*, 87-94.

Vogel, S., Ney, H., & Tillmann, C. (1996). HMM-based word alignment in statistical translation. In *Proceedings of the 16th conference on Computational linguistics*, 836-841.

Wu, D. (1997). Stochastic inversion transduction grammars and bilingual parsing of parallel corpora. *Computational Linguistics*, 23(3), 377-403.

Yamada, K. & Knight, K. (2001). A syntax-based statistical translation model. In *Proceedings of the 39th Annual Conference of ACL-2001*, 523-530.

Zens, R. & Ney, H. (2003). A comparative study on reordering constraints in statistical machine translation. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 144-151.

Zhang, H. & Gildea, D. (2004). Syntax-based alignment: supervised or unsupervised? In P*roceedings of the 20th International Conference on Computational Linguistics,* 418-424.

Zhang, H. & Gildea, D. (2005). Stochastic lexicalized inversion transduction grammar for alignment. In *Proceedings of the 43rd Annual Meeting of the ACL*, 475-482.

Zhang, H., Huang, L., Gildea, D., & Knight, K. (2006). Synchronous binarization for machine translation. In *Proceedings of the NAACL-HLT*, 256-263.

Computational Linguistics and Chinese Language Processing
Vol. 14, No. 1, March 2009, pp. 19-44

Automatic Sense Derivation for Determinative-Measure Compounds under the Framework of E-HowNet

Chia-Hung Tai*, Jia-Zen Fan*, Shu-Ling Huang*,+, and Keh-Jiann Chen*

Abstract

In this paper, we take Determinative-Measure Compounds as an example to demonstrate how the E-HowNet semantic composition mechanism works in deriving the sense representation for a newly coined determinative-measure (DM) compound. First, we define the sense of a closed set of each individual determiner and measure word in E-HowNet representation exhaustively. Afterwards, we make semantic composition rules to produce candidate sense representations for a newly coined DM. Then, we review development set to design sense disambiguation rules. We use these heuristic disambiguation rules to determine the appropriate context-dependent sense of a DM and its E-HowNet representation. The experiment shows that the current system reaches 89% accuracy in DM sense derivation and disambiguation.

Keywords: Semantic Composition, Determinative-Measure Compounds, Sense Representations, Extended How Net, How Net

1. Introduction

Building a knowledge base is time consuming work. The CKIP Chinese Lexical Knowledge Base has about 80 thousand lexical entries, and their senses are defined in terms of the E-HowNet format. E-HowNet is a lexical knowledge representation system. It extends the framework of HowNet (Dong *et al*., 2006) to allow semantic composition. Based on the framework of E-HowNet, we intend to establish an automatic semantic composition mechanism to derive sense of compounds and phrases from lexical senses (Chen *et al.,* 2005b),

* CKIP, Institute of Information Science, Academia Sinica
E-mail: {glaxy; kitajava; kchen} @iis.sinica.edu.tw
+ Department of Language and Literature Studies, National Hsinchu University of Education
E-mail: slhuang@mail.nhcue.edu.tw

[Received November 3, 2008; Revised May 27, 2009; Accepted May 28, 2009]

(Huang *et al.,* 2008). Determinative-Measure compounds (abbreviated as DM) are the most common compounds in Chinese. As a determiner and a measure normally coin a compound with unlimited versatility, the CKIP group does not define the E-HowNet representations for all DM compounds. Nevertheless, construction patterns for DMs are regular (Li *et al.,* 2006). Therefore, an automatic identification schema in regular expression (Li *et al.,* 2006) and a semantic composition method under the framework of E-HowNet for DM compounds were developed.

In this paper, we take DMs as an example to demonstrate how the E-HowNet semantic composition mechanism works in deriving the sense representations for all DM compounds. The remainder of this paper is organized as follows. Section 2 presents the background knowledge of DM compounds and sense representation in E-HowNet. We'll describe our method in Section 3 and discuss the experiment result in Section 4 before we present conclusions in Section 5.

2. Background

There are numerous studies on determiners as well as measures, especially on the types of measures[1]. Tai (1994) asserts that classifiers and measures words are often treated together under one single framework of analysis. Chao (1968) treats classifiers as one kind of measure word. In his definition, a measure is a bound morpheme which forms a DM compound with the determiners enumerated below.

i. Demonstrative determiners, *e.g.* 這 "this", 那 " that"…
ii. Specifying determiners, *e.g.* 每 "every", 各 "each"…
iii. Numeral determiners, *e.g.* 二 "two", 百分之三 "three percent", 四百五十 "four hundred and fifty"…
iv. Quantitative determiners, *e.g.* 一 "one", 滿 "full", 許多 "many"…

Measures are divided into nine classes by Chao (1968). Classifiers are defined as 'individual measures', which is one of the nine kinds of measures.

i. classifiers, *e.g.* 本 "a (book)",
ii. classifier associated with V-O constructions, *e.g.* 手 "hand",
iii. group measures, *e.g.* 對 "pair",
iv. partitive measures, *e.g.* 些 "some",

[1] Chao (1968) and Li and Thompson (1981) detect measures and classifiers. He (2002) traces the diachronic names of measures and mentions related literature on measures. The dictionary of measures pressed by Mandarin Daily News Association and CKIP (1997) lists all the possible measures in Mandarin Chinese.

v. container measures, *e.g.* 盒 "box",

vi. temporary measures, i. 身 "body",

vii. Standard measures, *e.g.* 公尺 "meter",

viii. quasi-measure, *e.g.* 國 "country",

ix. Measures with verb, *e.g.* 次 "number of times".

As mentioned in the introduction, Chao considered that determiners are listable and measures are largely listable, so D and M can be defined by enumeration, and that DM compounds have unlimited versatility. In this paper, we adopt the CKIP DM rule patterns and Part-of-Speeches for morpho-syntactic analysis, and, therefore, inherit the definition of determinative-measure compounds (DMs) in Mo *et al.* (1991). Mo *et al.* defined a DM as the composition of one or more determiners together with an optional measure. It is used to determine the reference or the quantity of the noun phrase that co-occurs with it. We use the definition of Mo *et al*. to apply to NLP and somewhat different from traditional linguistics definitions.

2.1 Regular Expression Approach for Identifying DMs

Due to the possible infinite number of DMs, Mo *et al.* (1991) and Li *et al.* (2006) proposed to identify DMs by regular expression as part of their morphological module in NLP. For example, when the DM compound is the composition of one determiner, *e.g.* numerals in (1), rules (2a), (2b), or (2c) will be first applied, and then rules (2d), (2e), or (2f) will be applied to compose complex numeral structures, and finally rule (2g) will generate the pos Neu of numeral structures. From the processes of regular expression, the numerals 534 and 319 in (1) are identified and tagged as Neu.[2]

(1) 鼓勵***534***人完成***319***鄉之旅

guli wubaisanshisi ren wancheng sanbaiyishijiu xiang zhi lu

encourage 534 persons to accomplish the travel around 319 villages

[2] The symbol "Neu" stands for Numeral Determiners. Generation rules for numerals are partially listed in (2).

(2) a. NO1 = {○,一,二,兩,三,四,五,六,七,八,九,十,廿,卅,百,千,
萬,億,兆,零,幾};
b. NO2 = {壹,貳,參,肆,伍,陸,柒,捌,玖,拾,佰,仟,萬,億,兆,零,幾};
c. NO3 = {１,２,３,４,５,６,７,８,９,０,百,千,萬,億,兆};
d. IN1 -> { NO1*, NO3*} ;
e. IN2 -> NO2* ;
f. IN3 -> {IN1,IN2} {多,餘,來,幾} ({萬,億,兆}) ;
g. Neu -> {IN1,IN2,IN3 } ;

Regular expression approach is also applied to deal with ordinal numbers, decimals, fractional numbers and DM compounds for times, locations etc.. The detailed regular expressions can be found in Li *et al.* (2006). Rule patterns in regular expression only provide a way to represent and to identify morphological structures of DM compounds, but do not derive the senses of complex DM compounds.

2.2 Lexical Sense Representation in E-HowNet

Core senses of natural language are compositions of relations and entities. Lexical senses are processing units for sense composition. Conventional linguistic theories classify words into content words and function words. Content words denote entities and function words mainly serve grammatical functions which link relations between entities/events. In E-HowNet, the senses of function words are represented by semantic roles/relations (Chen *et al.* 2005a). For example, 'because' is a function word. Its E-HowNet definition is shown in (3).

(3) because|因爲 def: reason={};

which means reason(x)={y} where x is the dependent head and y is the dependent daughter of '因爲'.

In the following sentence (4), we'll show how the lexical concepts are combined into the sense representation of the sentence.

(4) Because of the rain, all the clothes are wet. 因爲下雨，衣服都濕了

In the above sentence, ‘濕 wet’, ‘衣服 clothes’ and ‘下雨 rain’ are content words while ‘都 all’, ‘了 Le’ and ‘因爲 because’ are function words. The difference of their representation is that function words start with a relation but content words have under-specified relations. If a content word plays a dependent daughter of a head concept, the relation between the head concept and this content word will be established after parsing process. Suppose that the following dependent structure and semantic relations are derived after parsing the sentence (4).

(5) S(reason:VP(Head:Cb:因爲|dummy:VA:下雨)|theme:NP(Head:Na:衣服) |
quantity: Da:都 | Head:Vh:濕|particle:Ta:了)。

After the feature unification process, the following semantic composition result (6) is derived. The sense representations of dependent daughters became the feature attributes of the sentential head ‘wet|濕’.

(6) def:{wet|濕:
theme={clothing|衣物},
aspect={Vachieve|達成},
manner={complete|整},
reason={rain|下雨}}

In (5), the function word ‘因爲 (because)’ links the relation of ‘reason’ between head concept ‘濕 wet’ and ‘下雨 rain’. The result of the composition is expressed as reason(wet|濕)={rain|下雨}, since, for simplicity, the dependent head of a relation is normally omitted. Therefore, reason(wet|濕)={rain|下雨} is expressed as reason={rain|下雨}; theme(wet|濕)={clothing|衣物} is expressed as theme={clothing|衣物} and so on in the expression (6).

2.3 The sense representation for determiners and measures in E-HowNet

The sense of a DM compound is determined by its morphemes and the morphemes of DMs are determiners and measures which are exhaustively listable. Therefore, in order to apply a semantic composition mechanism to derive the senses of DM compounds, we first need to establish the sense representations for all determiners and measures. Determiners and measures are both modifiers of nouns/verbs and their semantic relation with head nouns/verbs are well established. We, thus, defined them by a semantic relation and its value like (7) and (8) below.

(7) The definition of determiners in E-HowNet

this 這 def: quantifier={definite|定指}

first 首 def: ordinal={1}

one 一 def: quantity={1}

For measure words, we found that some measure words contain content sense, but for some measure words, such as classifiers, their content senses are not important and could be neglected. So, we divided the measure words into two types, with or without content sense, with their sense representations being exemplified below:

(8) The definition of measure words in E-HowNet

a) Measure words with content sense

碗 bowl def: container={bowl|碗}

米 meter def: length={meter|公尺}

月 month def: time={month|月}

b) Measure words without content sense

本 copy def:{null}

間 room def:{null}

樣 kind def:{null}

3. Semantic Composition for DM Compounds

To derive sense representations for all DM compounds, we study how to combine the E-HowNet representations of determiners and measures into a DM compound representation and make rules for automatic composition accordingly. Basically, a DM compound is a composition of some optional determiners and an optional measure. It is used as a modifier to describe the quantity, frequency, container, length, etc. of an entity. The major semantic roles played by determiners and measures are listed in Table 1. Since an E-HowNet sense representation is basically a feature value structure, we will apply feature unification process for semantic derivation of DMs. The basic feature unification processes (Duchier *et al.,* 1999) is as follows:

If a morpheme *B* is a dependency daughter of morpheme *A*, *i.e*. *B* is a modifier or an argument of *A*, then unify the semantic representation of *A* and *B* via the following steps.

Step 1: Identify the semantic **relation** between *A* and *B* to derive **relation**(*A*)={*B*}.

Step 2: Unify the semantic representation of *A* and *B* by insert relation(*A*)={*B*} as a sub-feature of *A*.

As exemplified in (9) and (10), a feature unification process can derive the sense representation of a DM compound if its morpheme sense representations and semantic head are known.

(9) one 一 def:quantity={1} + bowl 碗 def: container={bowl|碗} →
one bowl 一碗 def: container={bowl|碗:quantity={1}}

(10) this 這 def: quantifier={definite|定指} + 本 copy def:{null} →
this copy 這本 def: quantifier={definite|定指}

Table 1. Major semantic roles played by determiners and measures

Semantic Role	D/M
quantifier	*e.g.* 這、那、此、該、本、貴、敝、其、某、諸
ordinal	*e.g.* 第、首
qualification	*e.g.* 上、下、前、後、頭、末、次、首、其他、其餘、別、旁、他、另、另外、各
quantity	*e.g.* 一、二、萬、雙、每、任何、一、全、滿、整、一切、若干、有的、一些、部份、有些、許多、很多、好多、好幾、好些、少許、許許多多、幾許、多數、少數、大多數、泰半、不少、個把、半數、諸多
Formal={.Ques.}	*e.g.* 何、啥、什麼
Quantity={over, approximate, exact}	*e.g.* 餘、許、足、之多、出頭、好幾、開外、整、正
position	*e.g.* 桌子、院子、地、屋子、池、腔、家子
container	*e.g.* 盒(子)、匣(子)、箱(子)、櫃子、櫥(子)、籃(子)、簍(子)、爐子、包(兒)、袋(兒)、池子、瓶(子)、桶(子)、聽、罐(子)、盆(子)、鍋(子)、籠(子)、盤(子)、碗、杯(子)、勺(子)、匙(湯匙)、筒(子)、擔(子)、籮筐、杓(子)、茶匙、壺、盅、筐、瓢、鍬、缸

length	*e.g.* 公厘、公分、公寸、公尺、公丈、公引、公里、市尺、營造尺、台尺、吋(inch)、呎(feet)、碼(yard)、哩(mile)、 (海)浬、庹、噚、尺、里、釐、寸、丈、米、厘、厘米、海 哩、英尺、英里、英呎、英寸、米突、米尺、微米、毫米、 英吋、英哩、光年
size	*e.g.* 公畝、公頃、市畝、營造畝、坪、畝、分、甲、頃、平方公里、平方公尺、平方公分、平方尺、平方英哩、英畝
weight	*e.g.* 公克、公斤、公噸、市斤、台兩、台斤(日斤)、盎司(斯)、磅、公擔、公衡、公兩、克拉、斤、兩、錢、噸、克、英磅、英兩、公錢、毫克、毫分、仟克、公毫
volume	*e.g.* 公撮、公升(市升)、營造升、台升(日升)、盎司、品脫(pint)、加侖(gallon)、蒲式耳(bushel)、公斗、公石、公秉、公合、公勺、斗、毫升、夸、夸特、夸爾、立方米、立方厘米、立方公分、立方公寸、立方公尺、立分公里、立方英尺、石、斛、西西
time	*e.g.* 微秒、釐秒、秒、秒鐘、分、分鐘、刻、刻鐘、點、點鐘、時、小時、更、夜、旬、紀(輪, 12 年) 、世紀、天(日)、星期(禮拜、週、周) 、月、月份、季、年(載、歲) 、週年、周歲、年份、晚、宿、世、輩、輩子、代、學期、學年、年代
address	*e.g.* 國、省、州、縣、鄉、村、鎮、鄰、里、郡、區、站、巷、弄、段、號、樓、衖、市、洲、地、街
place	*e.g* 部、司、課、院、科、系、級、股、室、廳
duration	*e.g* 陣(子)、會、會兒、下子

There are, however, some complications that must be resolved. First of all we have to clarify the dependent relation between the determiner and the measure of a DM in order to construct a correct feature unification process.

3.1 Head Morpheme of a DM Compound

In principle, a dependent head will take semantic representation of its dependent daughters as its features. Usually, determiners are modifiers of measures, such as '這 (this)' and '一 (one)' are modifiers of '碗 (bowl)' in the examples of 這碗, 一碗, 這一碗. For instance, Example (11) has the dependent relations of

NP(quantifier:DM(quantifier:Neu:一|container:Nfa:碗)|Head:Nab:麵)

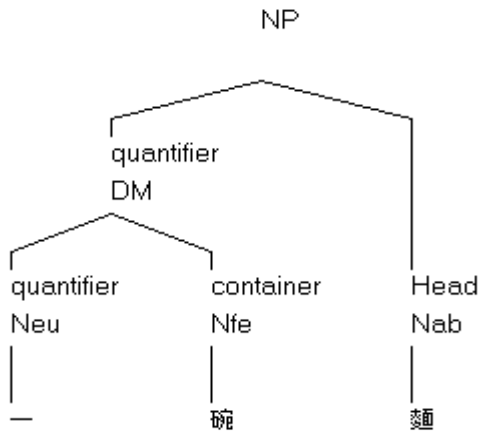

Figure 1. The dependent relations of 一碗麵 "a bowl of noddle".

After the feature unification process, the semantic representation of "一 def: quantity={1}" becomes the feature of its dependent head "碗 def: container={bowl|碗}" and derives the feature representation of "one bowl 一碗def: container={bowl|碗:quantity={1}}". Similarly, "one bowl 一碗" is the dependent daughter of "noodle|麵 def:{noodle|麵}". After the unification process, we derive the result of (11).

(11) one bowl of noodles|一碗麵 def:{noodle|麵:container={bowl|碗:quantity={1}}}

The above feature unification process, written in rule form, is expressed as (12).

(12) Determiner + Measure (D+M) → def: semantic-role(M) = {Sense-representation(M): Representation(D)}

The rule (12) says that the sense representation of a DM compound with a determiner D and a measure M is a unification of the feature representation of D as a feature of the sense representation of M as exemplified in (11).

Nevertheless, a DM compound with a null sense measure word, such as "this copy|這本" , "a copy|一本", or without measure word, such as "these three|這三", will be exceptions, since

the measure word cannot be the semantic head of DM compound. The dependent head of determiners become the head noun of the NP containing the DM and the sense representation of a DM is a coordinate conjunction of the feature representations of its morphemes of determiners only.

For instance, in (10), "copy" has weak content sense; therefore, we regard it as a null-sense measure word and only retain the feature representation of the determiner as the definition of "this copy|這本". The unification rule for DM with null-sense measure is expressed as (13).

(13) Determiner + {Null-sense Measure} (D+M) → def: Representation(D);

If a DM has more than one determiner, we can consider the consecutive determiners as one D and the feature representation of D is a coordinate conjunction of the features of all its determiners. For instance, "this one|這一" and "this one|這一本" both are expressed as "quantifier={definite|定指}, quantity={1}".

Omissions of numeral determiner occur very often while the numeral quantity is "1". For instance, "這本" in fact means "this one|這一本". Therefore, the definition of (10) should be modified as:

這本 def: quantifier={definite|定指}, quantity={1};

The following derivation rules cover the cases of omissions of numeral determiner.

(14) If both numeral and quantitative determiners do not occur in a DM,
then the feature quantity={1} is the default value of the DM.

Another major complication is that senses of morphemes are ambiguous. The feature unification process may produce many sense representations for a DM compound.

3.2 Sense Disambiguation

Multiple senses will be derived for a DM compound due to ambiguous senses of its morpheme components. For instance, the measure word "頭 (head)" has either the sense of {頭|head}, such as "滿頭白髮 full head of white hair" or the null sense in "一頭牛 a cow". Some DMs are inherent sense ambiguous and some are pseudo ambiguous. For instance, the above

example “一頭” is inherently ambiguous, since it could mean “full head” as in the example of “一頭白髮 full head of white hair” or could mean “one + classifier” as in the example of “一頭牛 a cow”. For inherently ambiguous DMs, the sense derivation step will produce ambiguous sense representations and leave the final sense disambiguation until seeing collocation context, in particular seeing dependent heads. Some ambiguous representations are improbable sense combination. The improbable sense combinations should be eliminated during or after feature unification of D and M. For instance, although the determiner “一” has ambiguous senses of “one”, “first”, and “whole”, “一公尺” has only the sense of “one meter”, so the other sense combinations should be eliminated.

The way we tackle the problem is that first we find all the ambiguous Ds and Ms by looking their definitions shown in Appendix A. We, then, manually design content and context dependent rules to eliminate the improbable combinations for each ambiguous D or M types. For instance, according to Appendix A, “頭” has 3 different E-HowNet representations while it functions as a determiner or measure, *i.e.* “def:{null}”, “def:{head|頭}”, and “def:ordinal={1}”. We write three content or context dependent rules below to disambiguate its senses.

(15) 頭 “head”, Nfa, E-HowNet: “def:{null}” : while E-HowNet of the head word is “動物({animate|生物})” and its subclasses.

(16) 頭 “head“, Nff, E-HowNet: “def:{head|頭}” : while pre-determiner is 一(Neqa) “one” or 滿 “full” or 全 “all” or 整 “total”.

(17) 頭 “first”, Nes, E-HowNet: “def:ordinal={1}” : while this word is being a demonstrative determiner that is a leading morpheme of the compound.

The disambiguation rules are shown in Appendix B. In each rule, the first part is the word and its part-of-speech. Then, the E-HowNet definition of this sense is shown, followed by the condition constraints for this sense. If there is still remaining ambiguity after using the disambiguation rule, we choose the most frequent sense as the result.

3.3 Simplification and Normalization for Sense Representation

Members of every type of determiners and measures are exhaustively listable except numeral determiners. Also, the formats of numerals are various. For example, "5020" is equal to "五零二零" and "五千零二十" and "五千二十". So, we have to unify the numeral representation into a standard form. All numerals are composed of basic numerals, as shown in the regular expressions (2). Their senses, however, are not possible to define one by one. We take a simple approach. For all numerals, their E-HowNet sense representations are expressed as themselves. For example, 5020 is expressed as quantity={5020} and we will not further define the sense of 5020. Furthermore all non-Arabic forms will be converted into Arabic expressions, *e.g.* "五千零二十" is defined as quantity={5020}.

The other problem is that the morphological structures of some DMs are not regular patterns. Take "兩個半(two and a half)" as an example. "半(half)" is not a measure word. So, we collect those words, like "多 (many), 半 (half), 幾 (many), 上 (up), 大 (big), 來 (more)" to modify the quantity definition. So, we first remove the word "半" and define the "兩個" as quantity={2}. As the word "半" means quantity={0.5}, we define the E-HowNet definition for "兩個半" as quantity={2.5}. For other modifiers such as "多 (many), 幾 (many), 餘 (more), 來 (more)," we use a function over() to represent the sense of "more", such as "十多個 more than 10" is represented as quantity={over(10)}.

In E-HowNet, complex word senses are expressed by some limit number of basic or primitive concepts. Nevertheless, some certain domain concepts can hardly be expressed by primitive concepts, for instance "焦耳 (joule)," "盧比 (rupee)," "五千零二十 (five thousand and twenty)," etc.. Therefore, we simplify our representations and consider many domain specific concepts as basic concept without further decomposing into primitive concepts.

Appendix A shows the determiners and measures used and their E-HowNet definition in our method. Now, we have the basic principles for compositing semantics of DM under the framework of E-HowNet.

The following steps show how we process DMs and derive their E-HowNet definitions from an input sentence.

I. Input: a Chinese sentence.
II. Apply regular expression rules for DM to identify all possible DM candidates in the input sentence.
III. Segment DM into a sequence of determiners and measures.
IV. Normalize numerals into Arabic form if necessary
V. Apply feature unification rules (12-14) to derive candidates of E-HowNet representations for every DM.

VI. Disambiguate candidates for each DM if necessary.

VII. Output: DM Compounds in E-HowNet representation.

For an input Chinese sentence, we use the regular expression rules created by Li *et al.* (2006) to identify all possible DMs in the input sentence. Then, for every DM compound, we segment it into a sequence of determiners and measures. If any numerals exists in the DM, every numeral is converted into decimal number in Arabic form. For every DM, we follow the feature unification principles to composite semantics of DM in E-HowNet representations and produce possible ambiguous candidates. Then, the final step of sense disambiguation will be carried out.

4. Experiments and Discussion

A corpus-based approach was adopted in developing our proposed method. We need a developing set to derive an exhaustive list of determiners and measures. We try to extract DMs and their morpheme components, *i.e.* determiners and measures, from the developing set and observe the instances of DM to decide their senses and sense representations. Furthermore, sense disambiguation rules will also be developed according to the context of sense ambiguous instances. First, we need to know how many DMs are sufficient to derive a list of determiners and measures with high coverage, if it is not exhaustive. Therefore, we extract DMs from different size subsets of Sinica Treebank and observe their character token coverage. The results are shown in Table 2 and Figure 2. We find that the set of determiners and measures extracted from more than 15000 sentences is sufficient to cover more than 99% of DM instances in the Sinica Treebank.

Table 2. The character token coverage of different subsets of Sinica Treebank

Sentences	0	5000	10000	15000	20000	25000	30000
DM char distribution coverage	0	0.971816	0.987363	0.994014	0.996259	0.997755	0.999169

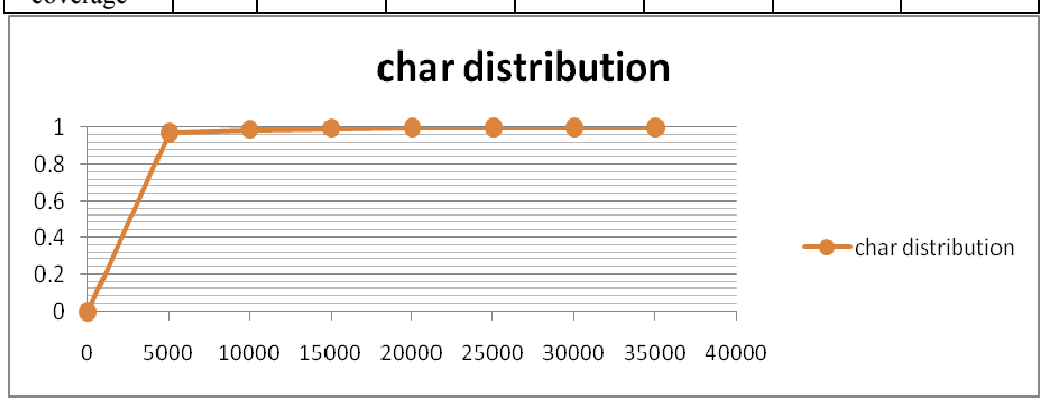

Figure 2. The growth diagram of DM character token coverage.

Therefore, we randomly selected 16070 sentences from Sinica Treebank as our development set and 10000 sentences as our testing set. The development set contained 3753 DM tokens and the testing set contained 1604 DM tokens. We used the development set to derive lexical sense representations and to design disambiguation rules. A total of 405 determiner types and 211 measure types were found, in which 367 out of the 405 determiners were numerals. Since the numbers of numeral determiners are infinite, all numerals will be converted into their Arabic form automatically instead of representing their E-HowNet sense representations individually. The rest of the determiners and measures are encoded with their E-HowNet sense representations manually. For words with ambiguous senses, we also derived their disambiguation rules according to their contextual information shown in development corpus. Finally, a total of 40 disambiguation rules were developed, as shown in Appendix B.

The sense representations of a DM compound will then be derived by a semantic composition process under the framework of E-HowNet. The evaluation of the sense derivation for DM compounds can be divided into two parts: the first part is the correctness of the semantic composition process, and the second part is the correctness of the sense disambiguation process.

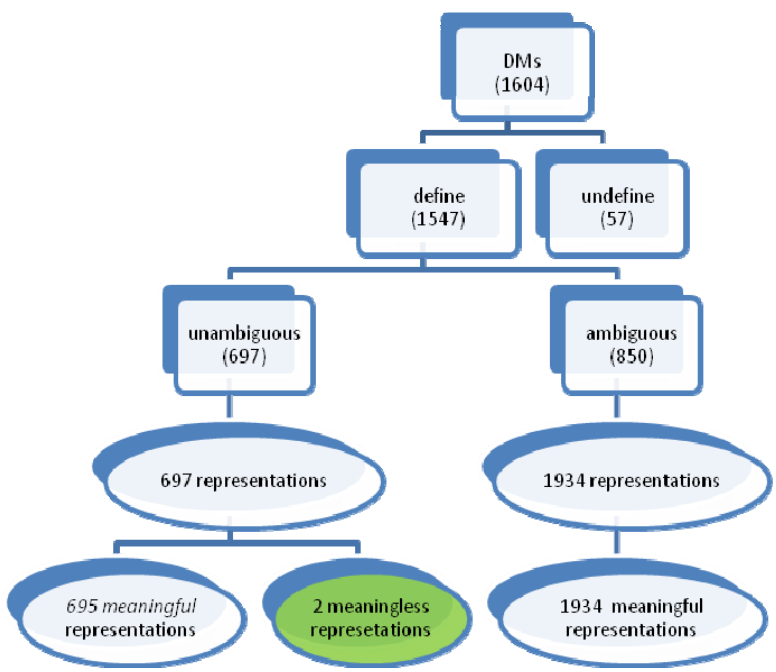

Figure 3. The evaluation result of the semantic composition process.

Figure 3 shows the evaluation result of the semantic composition process. The semantic composition process produced 2631 representations from 1604 words. The program failed to produce E-HowNet representations for the remaining 57 words because of undefined

morphemes. Ambiguous senses were found in 850 words out of the 1604 words. The quality of the result candidates is pretty good. Table 3 shows some sample results. For testing the correctness of our candidates, we checked the formats of 2631 candidates manually. Only 2 candidates out of 2631 displayed wrong or meaningless representations, with both coming from unambiguous words. Therefore, the covering ratio of semantic composition process, *i.e.* deriving meaningful representation without considering sense correctness, is 96% ((1547-2)/1604).

Table 3. Sample results of semantic composition for DM compounds.

DM Compounds	E-HowNet Representation
二十萬元	def:role={money|貨幣:quantity={200000}}
另一個	def:qualification={other|另},quantity={1}
二百三十六分	def:role={score|分數:quantity={236}}
前五天	def:time={day|日:qualification={preceding|上次}, quantity={5}}
一百一十六點七億美元	def:role={USD|美元:quantity={11670000000}}

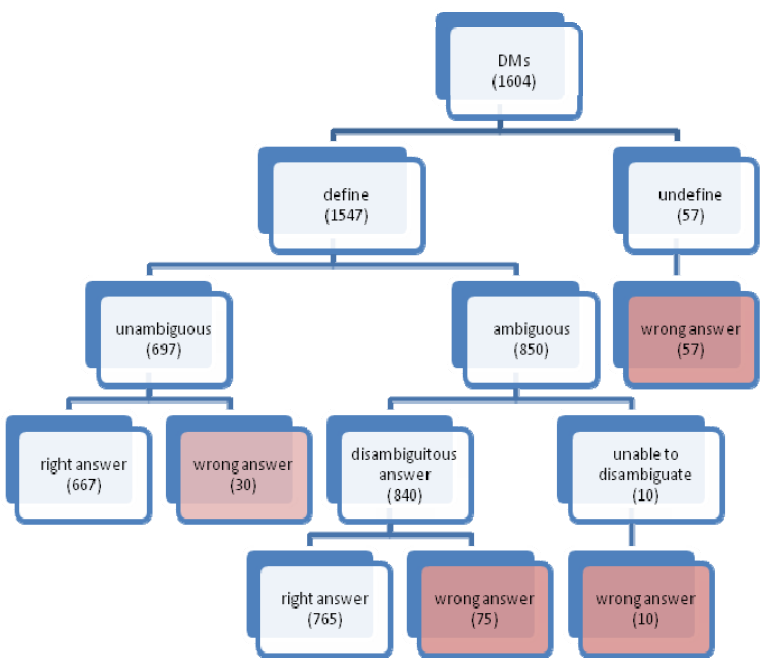

Figure 4. The accuracy of composed sense for DM compounds.

For checking sense correctness, after the disambiguation processes, the resulting E-HowNet representations of 1604 DM tokens in their context were judged manually. Among them, 850 token DMs were sense-ambiguous and the composition process failed to generate answers for 10 of them. Therefore, the composition rules cover 98.8% (840/850) of the ambiguous DM tokens and the precision of the disambiguation rules is 91% (765/840). In all, there are 1432 correct E-HowNet representations for 1604 DM tokens in both sense and format, *i.e*. the current model achieves 89% ((667+765)/1604) token accuracy. Among the 172 wrong answers, 57 errors are due to undefined ambiguous morpheme sense, 30 errors are unique but the wrong answer, and there are 85 sense disambiguation errors.

After data analysis, we conclude the following error types.

A. Unknown domain error:

七棒 "7th batter", 七局 "7th inning"

As there is no text related to the baseball domain in the development set, we get poor performance in dealing with text about baseball. The way to resolve this problem is to increase the coverage of sense representations and disambiguation rules for the baseball domain.

B. Sense ambiguities:

In the following parsed phrase, NP(property:DM:上半場 "first half "|Head:DM:二十分 "twenty minutes or twenty points"), the E-HowNet representation of 二十分 "twenty minutes or twenty points" can be defined as "def:role={score|分數:quantity={20}}" or "def:time={minute|分鐘:quantity={20}}". More contextual information is required to resolve such kinds of sense ambiguity.

For the type of unknown domain error, the solution is to expand the disambiguation rules and the sense representations for morphemes. For sense ambiguities, we need more information and better features to determine true senses.

5. Conclusion

E-HowNet is a lexical sense representational framework and intends to achieve sense representation for all compounds, phrases, and sentences through automatic semantic composition processing. For this purpose, we defined word senses of the CKIP Chinese lexicon in E-HowNet representation. Then, we tried to automate semantic composition for phrases and sentences. Nevertheless, many unknown words or newly coined compound words may occur in the target sentences. In fact, DM compounds are the most frequently occurring unknown words. Therefore, our first goal was to derive the senses of DM words automatically.

In this paper, we take DMs as an example to demonstrate how the semantic composition mechanism works in E-HowNet to derive the sense representations for all DM compounds. We analyze morphological structures of DMs and derive their morphological rules in terms of regular expression. Then, we defined the sense of all determiners and measures in E-HowNet format exhaustively. We created some simple composition rules to produce candidate sense representations for DMs. Then, we reviewed the development set to write some disambiguation rules. We used these heuristic rules to determine the final E-HowNet representation and reach 89% accuracy. The current version did not exhaustively collect all determiners and measures. The system, however, can be improved by gradual extension of the representations of new determiners and measures without retraining.

In the future, we will use similar methods to handle general compounds and to improve sense disambiguation and semantic relation identification processing. We intend to achieve semantic compositions for phrases and sentences in the future and we had shown the potential in this paper.

Acknowledgement

This research was supported in part by the National Science Council under a Center Excellence Grant NSC 96-2752-E-001-001-PAE and Grant NSC96-2221-E-001-009.

Reference

Chao, Y. R. (1968). *A grammar of Spoken Chinese,* University of California Press, Berkeley.

Chen, K. J., Huang, S. L., Shih, Y. Y., & Chen, Y. J. (2005a). Extended-HowNet- A Representational Framework for Concepts. In *Processing of OntoLex 2005 - Ontologies and Lexical Resources IJCNLP-05 Workshop*, Jeju Island, South Korea.

Dong, Z. D. & Dong, Q. (2006). *HowNet and the Computation of Meaning*, World Scientific Publishing Co. Pte. Ltd.

Duchier, D., Gardent, C., & Niehren, J. (1999). Concurrent constraint programming in Oz for natural language processing. Lecture notes, http://www.ps.uni-sb.de/~niehren/oz-natural-language-script.html.

Huang, S. L., Chung, Y. S., & Chen, K. J. (2008). E-HowNet- an Expansion of HowNet. In *Proceedings of 1st National HowNet Workshop*, Beijing, China.

Li, C. N., & Thompson, S. A. (1981). *Mandarin Chinese: A Functional Reference Grammar*, University of California Press, Berkeley.

Li, S. M., Lin, S. C., Tai, C. H., & Chen, K. J. (2006). A Probe into Ambiguities of Determinative-Measure Compounds. *International Journal of Computational Linguistics & Chinese Language Processing*, 11(3), 245-280.

Mo, R. P., Yang, Y. J., Chen, K. J., & Huang, C. R. (1991). Determinative-Measure Compounds in Mandarin Chinese: Their Formation Rules and Parser Implementation. In

Proceedings of ROCLING IV (R.O.C. Computational Linguistics Conference), National Chiao-Tung University, Hsinchu, Taiwan, 111-134.

Tai, J. H.Y. (1994). Chinese classifier systems and human categorization. *In Honor of William S.Y. Wang: Interdisciplinary Studies on Language and Language Change*, ed. by M. Y. Chen and O. J. L. Tzeng, Pyramid Press, Taipei, 479-494.

何杰(He, J.), (2002), *現代漢語量詞研究*, 民族出版社, 北京市。

陳怡君(Chen, Y. C.), (2005b), 黃淑齡, 施悅音, 陳克健, "繁體字知網架構下之功能詞表達初探", 收錄於*第六屆漢語詞彙語意學研討會論文集*, 廈門大學, 中國。

黃居仁(Huang, C. R.), (1997), 陳克健, 賴慶雄(編著), *國語日報量詞典*, 國語日報出版社, 台北。

Appendix A. Determiner and measure words in E-HowNet representation

定詞(Determiners)

定指

D1->這、那、此、該、本、貴、敝、其、某、諸 def: quantifier={definite|定指}；這些、那些 def: quantifier={definite|定指}, quantity={some|些}

D2->第、首 def: ordinal={D4}

D3->上、前 def: qualification={preceding|上次}；下、後 def: qualification={next|下次}；頭、首 def:ordinal={1}；末 def: qualification={last|最後}；次 def:ordinal={2}

不定指

D4->一、二、萬、雙... def: quantity={1、2、10000、2...} or def:ordinal={1、2、10000、2...}

D5->甲、乙... def: ordinal={1、2...}

D6->其他、其它、其餘、別、旁、他、另、另外 def: qualification={other|另}

D7->每、任何、一、全、滿、整、一切 def: quantity={all|全}

D8->各 def: qualification={individual|分別的}

D9->若干、有的、一些、部份、有些、部分、些 def: quantity={some|些}

D10->半 def: quantity={half|半}

D11->多少、幾多、若干 def: quantity={.Ques.}

D12->何、啥、什麼 def: fomal={.Ques.}

D13->數、許多、很多、好多、好幾、好些、多、許許多多、多數、大多數、不少、泰半、半數、諸多 def: quantity={many|多}；少許、少數、幾許、個把 def: quantity={few|少}；幾 def:quantity={some|些}

D14->餘、許、之多、來 def: approximate()；足、整、正 def: exact()；出頭、數、好幾、幾、開外、多 def: over();

D15->0、1、2、3、4、5、6、7、8、9、０、１、２、３、４、５、６、７、８、９ def: quantity={1、2、3、4...}

量詞(Measure word)

有語意量詞(Measures with content sense)

Nff->暫時量詞—身、頭、臉、鼻子、嘴、肚子、手、腳 def:{身,頭, …}

Nff->暫時量詞—桌、桌子、院子、地、屋子、池、腔、家子 def: position={桌子,院子...:quantity={all|全}}

Nfe->容器量詞—

盒(子)、匣(子)、箱(子)、櫃(子)、櫥(子)、籃(子)、簍(子)、爐(子)、包(兒)、袋(兒)、池子、瓶(子)、桶(子)、罐(子)、盆(子)、鍋(子)、籠(子)、盤(子)、碗、杯(子)、勺(子)、匙(湯匙)、筒(子)、擔(子)、籮筐、杓(子)、茶匙、壺、盅、筐、瓢、鍬、缸 def: container={盒,匣,...}

Nfg->標準量詞—

表長度的，如：公厘、公分、公寸、公尺、公丈、公引、公里、市尺、營造尺、台尺、吋(inch)、呎(feet)、碼(yard)、哩(mile)、(海)浬、海里、庹、噚、尺、里、釐、寸、丈、米、厘、厘米、海哩、英尺、英里、英呎、英寸、米突、米尺、微米、毫米、英吋、英哩、光年。 def: length={公分,...}

表面積的，如：公畝、公頃、市畝、營造畝、坪、畝、分、甲、頃、平方公里、平方公尺、平方公分、平方尺、平方英哩、英畝。def: size={公畝,...}

表重量的，如：公克、公斤、公噸、市斤、台兩、台斤(日斤)、盎司(斯)、磅、公擔、公衡、公兩、克拉、斤、兩、錢、噸、克、英磅、英兩、公錢、毫克、毫分、仟克、公毫。def: weight={公克,...}

表容量的，如：公撮、公升(市升)、營造升、台升(日升)、盎司、品脫(pint)、加侖(gallon)、蒲式耳(bushel)、公斗、公石、公秉、公合、公勺、斗、毫升、夸、夸特、夸爾、立方米、立方厘米、立方公分、立方公寸、立方公尺、立分公里、立方英尺、石、斛、西西。def: volume={公撮,公升,...}

表時間的，如：微秒、釐秒、刻、刻鐘、點、點鐘、更、旬、紀(輪, 12年) 、世紀、季 def:time={微秒,釐秒,…}；秒、秒鐘 def:time={second|秒}；分、分鐘 def:time={minute|分鐘}；時、小時 def:time={hour|時}；夜、晚、宿 def:time={night|夜}；天(日) def:time={day|日}；星期(禮拜、週、周) def:time={week|周}；月、月份 def:time={month|月}；年、載、歲、年份 def:time={year|年}；週年、周歲 def:duration={年}

表錢幣的，如：元(圓)、塊、兩 def:role={money|貨幣}；分、角(毛)、先令、盧比、法郎(朗)、辨士、馬克、鎊、盧布、美元、美金、便士、里拉、日元、日圓、台幣、港幣、人民幣。def: role={分, …,盧布…}

其他：刀、打(dozen)、令、綸(十條)、蘿(gross)、大籮(great gross)、焦耳、千卡、仟卡、燭光、千瓦、仟瓦、伏特、馬力、爾格(erg)、瓦特、瓦、卡路里、卡、仟赫、位元、莫耳、毫巴、千赫、歐姆、達因、兆赫、法拉第、牛頓、赫、安培、周波、赫茲、分貝、毫安培、居里、微居里、毫居里 def: quantity={刀,打,…,焦耳,...}

Nfh->準量詞—

指行政方面，如：部、司、課、院、科、系、級、股、室、廳。def: location={部,

司...}

指時間方面，如：世、輩、輩子、代、學期、學年、年代 def: time={學期, 年代,...} 會、會兒、陣(子) 、下(子) def: duration={TimeShort|短時間}

指方向的，如：面(兒)、方面、邊(兒)、方 def: direction={EndPosition|端}；頭(兒) def: direction={aspect|側}

指音樂的，如：拍、小節。def: quantity={拍,板...}

指分數，如：分 def:role={分數:quantity={D4,D15}}；

Nfi->動量詞—

指頻率的，如：回、次、遍、趟、下、巡、遭、響、圈、把、關、腳、巴掌、掌、拳頭、拳、眼、口、刀、槌(子)、板(子)、鞭(子)、棒、棍(子)、針、槍矛、槍、砲、度、輪、周、跤、回合、票 def:frequency={D4, D15}；步 def:{步}；箭 def:role={箭:quantity={D4,D15}}；曲 def:{曲:quantity={D4,D15}}

Nfc->群體量詞—

對、雙 def:quantity={double|複}；

列(系列)、排 def:quantity={mass|眾:manner={ InSequence|有序}}；

套 def:quantity={mass|眾:manner={relevant|相關}}；

串 def:quantity={mass|眾:dimension={linear|線}}；

掛、幫、群、伙(夥)、票、批 def: quantity={mass|眾}；

組 def: quantity={mass|眾: manner={relevant|相關}}；

窩 def: quantity={mass|眾:cause={assemble|聚集}}；

種、類、樣 def: {kind({object|物體})}；

簇 def:quantity={mass|眾:cause={assemble|聚集}}；

疊 def:quantity={mass|眾:cause={pile|堆放}}；

紮 def:quantity={mass|眾:cause={wrap|包紮}}；

叢 def:quantity={mass|眾:cause={assemble|聚集}}；

隊 def:quantity={mass|眾:manner={InSequence|有序}}；

式 def:{kind({object|物體})}

Nfd->部分量詞—

些 def:quantity={some|些}；

部分(份)、泡、綹、撮、股、灘、汪、帶、截、節 def: quantity={fragment|部}；

團 def: quantity={fragment|部:shape={round|圓}}；

堆 def: quantity={ fragment|部:cause={pile|堆放}}；

把 def: quantity={ fragment|部:cause={hold|拿}}；

層、重 def: quantity={ fragment|部:shape={layered|疊}}

Nfa->個體量詞

號 def:ordinal={}

無語意量詞(null-sense Measures)

Nfa->個體量詞—洞、號、渠、本、把、瓣、部、柄、床、處、期、齣、場、朵、頂、堵、道、頓、錠、棟(幢)、檔(檔子)、封、幅、發、分(份)、人份、紙、服、個(箇)、根、行、戶、件、家、架、卷、具、闕、句、屆、捲、劑、隻、尊、盞、張、枝(支)、樁、幀、只、株、折、炷、軸、口、棵、款、客、輛、粒、輪、枚、面、門、幕、匹、篇、片、所、艘、扇、首、乘、襲、頭、條、台、挺、堂、帖、顆、座、則、冊、任、尾、味、位、頁、葉、房、彎、班、員、科、丸、名、項、起、間、題、目、招、股、回、線、灣。def: {null}

Nfc->群體量詞—宗、畦、餐、行、副(付)、蓬、筆、房、綑(捆)、胎、嘟嚕、部、派、路、壟、落、束、席、色、攤、項、疊、紮。def: {null}

Nfd->部分量詞—口、塊、滴、欄、捧、抱、段、絲、點、片、縷、坨、匹、疋、階、抔、波、道。def: {null}

Nfb->述賓式合用的量詞—通、口、頓、盤、局、番。def: {null}

Nfi->動量詞—回、次、遍、趟、下、遭、聲、響、圈、把、仗、覺、頓、關、手、(巴)掌、拳(頭)、眼、口、槌(子)、板(子)、鞭(子)、棒、棍(子)、針、箭、槍(矛)、砲、度、輪、曲、跤、記、回合、巡、票。def: {null}

Nfh->準量詞

指書籍方面，如：版、冊、編、回、章、面、小節、集、卷。def: {null}

指筆劃方面，如：筆、劃(兒)、橫、豎、直、撇、捺、挑、剔、鉤(兒)、拐、點、格(兒)。def: {null}

其他：

程、作(例:一年有兩作)、倍、成。def: {null}

厘(例:年利五厘、一分一厘都不能錯)。def: {null}

毫(萬分之一)、絲(十萬分之一)(例:一絲一毫都不差)。

圍、指、象限、度。def: {null}

開(指開金)、聯(例:上下聯不對稱)。def: {null}

軍、師、旅、團、營、伍、班、排、連、球、波、端。def: {null}
樓、城(扳回一城)、回合、折、摺、流、等、桿、聲、次。def: {null}

Appendix B. The rules for candidate disambiguation

Head-Based Rules

Rule 1 一, Neu, def:quantity={1}, while part-of-speech of the head word is Na, except the measure word is 身 "body" or 臉 "face" or 鼻子 "nose" or 嘴 "mouth" or 肚子 "belly" or 腔 "cavity" .

Rule 2 塊, Nfg, def:role={money|貨幣}, while E-HowNet representation of the head word is "{money|貨幣}" or {null}, or the head word is 錢 "money" or 美金 "USD" or the suffix of word is 幣 "currency" and previous word is not D1.

Rule 3 塊, Nfd, def:{null}, otherwise, use this definition.

Rule 4 面,Nfa,def:{null}, while part-of-speech of the head word is Nab.

Rule 5 面,Nfh,def:direction={aspect|側}, otherwise use this one.

Rule 6 頭,Nfa,def:{null}, while the head word is Nab and E-HowNet representation of the head word is "動物{animate|生物}".

Rule 7 頭,Nfh,def:direction={EndPosition|端}, if the part-of-speech of the head word is Na, do not use this definition. The prefix determiners are 這 "this" or 那 "that" or 另 "another".

Rule 8 All Nfi, def:frequency={}, while the part-of-speech of the head word is Verb, i.e. E-HowNet representation of the head word is {event|事件} and it's subclass. Except POS of the head are V_2 and VG, and if the word is {次、回、口}, do not use this rule.

Rule 9 All Nfi, def:{null}, otherwise use this one. If the head word is {次、回、口}, do not use this rule.

Rule 10 部, 股…,Nfh,def:location={ }, if part-of-speech of the head word is Na or prefix determiner is 這"this" or 那"that" or 每"every", do not use this definition.

Rule 11 部,股…,Nfa,def:{null}, otherwise use this definition.

Rule 12 盤,Nfe,def:container={plate|盤},while the head word is food, *i.e*. E-HowNet representation of the head word is {edible|食物} and its subclasses.

Rule 13 盤,Nfb,def:{null},otherwise use this one.

Rule 14 分,Nfg, def:role={分}, while the head word is 錢 "money", *i.e*. E-HowNet representation of the head word is {money|貨幣} and its subclasses.

Rule15 分,Nfg, def:size={分}, while the head word is 地 "land", *i.e*. E-HowNet representation of the head word is {land|陸地} and its subclasses.

Rule 16 分,Nfa,def:{null}, while part-of-speech of the head word is Na or Nv. For example: 一分耕耘；十分力氣；五分熟.

Rule 17 點,Nfh;Nfd,def:{null}, while part-of-speech of the head word is Nab. If part-of-speech of the head word is V, Naa or Nad, do not use this definition.

Rule 18 度,聲, def:frequency={}, while part-of-speech of the head word is Verb.

Rule 19 度,聲, def:{null}, otherwise use this definition.

Collocation-Based Rules

Rule 20 分,Nfh,def:role={score|分數:quantity={D4,D15}}, while the sentence also contains the words 考 "give an exam" (E-HowNet representation is {exam|考試}) or 得 "get" (E-HowNet representation is {obtain|得到}) or 失 "lose" (E-HowNet representation is {lose|失去}) or E-HowNet representation of the head word is {hold|拿},{catch|捉住},{occupy|佔領},{rob|搶},{win|獲勝},{forming|形成},{add|增加},{suffer|遭受},{sink|下沉},{inferior|不如} and its subclasses, or the sentence contains the word 成績 "score", X 局 (for example,一局 "the first inning"), X 半場 (for example, 上半場 "the first half in game"), then use this definition.

Rule 21 分,Nfg,def:time={minute|分鐘}, if the sentence contains the word 時 "hour", 鐘頭 "hour", X 時(for example,五時 "5 o'clock") or X 秒 (for example,三十秒 "30 seconds").

Rule 22 兩,Nfg,def:weight={兩}, if the sentence contains the word 重 "weight" or 重量 "weight".

Rule 23 兩,Nfg,def:role={money|貨幣}, if the sentence contains the word 銀 "silver" or 錢 "money" or 黃金 "gold"

Pre-Determinant-Based Rule

Rule 24 頭, Nff,def:{head|頭}, while the pre-determinant is 一(Neqa) "one" or 滿 "full" or 全 "all" or 整 "total".

Rule 25 腳, Nff,def:{leg|腳}, while pre-determinant is 一(Neqa) "one" or 滿 "full" or 全 "all" or 整 "total" and the part-of-speech of the head word is not Na.

Rule 26 腳, Nfi,def:frequency={}, while part-of-speech combination is V+D4,D15+腳.

Rule 27 點,Nfg, def:time={點}, while part-of-speech of pre-determiners are D4 or D15(1~24) and part-of-speech of the previous word is not D1 or the previous word is not 有 "have".

Determinative-Based Rule

Rule 28 一、二...1、2...兩..., Neu, def:ordinal={}, the determiners are 第, 民國, 公元, 西元, 年號, 一九 XX or 12XX, (four digits number).

Rule 29 一、二...1、2...兩..., Neu,def:quantity={}, otherwise use this definition.

Rule 30 頭,Nes,def:ordinal={1},the word 頭 “head” is a determiner.

Rule 31 兩,Neu,def:quantity={}, the word 兩 “a unit of weight equal to 50 grams” is a determiner.

Measure Word-Based Rule

Rule 32 一,Neqa,def:quantity={all|全}, the part-of-speech of the measure word behind 一 is Nff, or the suffix of the measure word is 子, (for example, 櫃子 “cabinet”, 瓶子 “bottle”) or 籮筐 “large basket”.

Rule 33 一、二...1、2...兩..., Neu,def:ordinal={}, if measure word is 歲.

Head and Determinantive-Based Rule

Rule 34 次,Nfi,def:frequency={}, while part-of-speech of the head word is a Verb (Except POS V_2 and VG.), and determiners are not D1,D2,D3.

Rule 35 次,Nfi;Nfh,def:{null}, otherwise use this definition.

Rule 36 口,Nfa.def:{口:quantity={全}}, while the pre-determiners are 滿,全,or 整.

Rule 37 口,Nfi.def:frequency={}, while part-of-speech of the head word is Verb, and the pre-determiner is not 滿,全,or 整.

Rule 38 口,def:{null}, otherwise, while the pre-determiner is D4 or D15, use this definition.

Rule 39 回, def:frequency={}, while part-of-speech of the head word is Verb (Except POS V_2 and VG), and the determiner is not D1,D2,D3.

Rule 40 回, def:{null}, otherwise use this definition.

Computational Linguistics and Chinese Language Processing
Vol. 14, No. 1, March 2009, pp. 45-84

Assessing Text Readability Using Hierarchical Lexical Relations Retrieved from WordNet

Shu-Yen Lin*, Cheng-Chao Su*, Yu-Da Lai*, Li-Chin Yang* and Shu-Kai Hsieh*

Abstract

Although some traditional readability formulas have shown high predictive validity in the r = 0.8 range and above (Chall & Dale, 1995), they are generally not based on genuine linguistic processing factors, but on statistical correlations (Crossley *et al.*, 2008). Improvement of readability assessment should focus on finding variables that truly represent the comprehensibility of text as well as the indices that accurately measure the correlations. In this study, we explore the hierarchical relations between lexical items based on the conceptual categories advanced from Prototype Theory (Rosch *et al.*, 1976). According to this theory and its development, basic level words like *guitar* represent the objects humans interact with most readily. They are acquired by children earlier than their superordinate words like *stringed instrument* and their subordinate words like *acoustic guitar*. Accordingly, the readability of a text is presumably associated with the ratio of basic level words it contains. WordNet (Fellbaum, 1998), a network of meaningfully related words, provides the best online open source database for studying such lexical relations. Our study shows that a basic level noun can be identified by its ratio of forming compounds (*e.g.* chair → armchair) and the length difference in relation to its hyponyms. We compared graded readings for American children and high school English readings for Taiwanese students by several readability formulas and in terms of basic level noun ratios (*i.e.* the number of basic level noun types divided by the number of noun types in a text). It is suggested that basic level noun ratios provide a robust and meaningful index of lexical complexity, which is directly associated with text readability.

* Department of English, National Taiwan Normal University
E-mail: {yenyenet, vennysu, yudalai, lchyang1112, shukai}@gmail.com

[Received January 1, 2009; Revised April 11, 2009; Accepted April 15, 2009]

Keywords: Readability, Prototype Theory, WordNet, Basic Level Words, Compounds.

1. Introduction

Traditional methods of measuring text readability typically rely on surface-level linguistic information such as the counting of sentences, words, syllables, or letters. Caution has long been taken in correlating these formulas with the reading process (Davison & Kantor, 1982; Rubin, 1985). In light of the many psycholinguistic findings on the reading process (Just & Carpenter, 1987; Perfetti, 1985; Rayner & Pollatsek, 1994), we start our research by assuming, in line with Rosch *et al.*'s Prototype Theory (Rosch & Mervis, 1975, Rosch *et al.*, 1976) and its later development (Rosch, 1977, 1978; Coleman & Kay, 1981; Lakoff, 1986; Tversky, 1990; Ungerer & Schmid, 1996), that words form conceptual hierarchies (*e.g.* furniture → chair → armchair) with lexical items at different levels posing varied processing difficulties. Putting the logic into templates, the measurement of the lexical difficulty of a text may be done by calculating the hierarchical levels at which its words fall. The best tool for our study is WordNet, a large, open source electronic lexical database of English, in which the different senses of words are interlinked in hierarchical structures by means of conceptual-semantic relations.

Our research was comprised of two stages. In the preliminary experiments, we utilized WordNet to identify the characteristics of basic level nouns. It was found that a basic level noun can be identified by its ratio of forming compounds (*e.g.* chair → armchair) and the length difference in relation to its full hyponyms. In the subsequent experiment, we compared selected readings in terms of their basic level noun ratios and their values calculated by several readability formulas. It is shown that basic level noun ratios are highly correlated with the text levels. Our study also indicates that there is a basic level in a lexical hierarchy which is easier to comprehend than its upper or lower levels. This finding challenges the intuitive idea underlying McNamara *et al.* (2002) that a word having more hypernym levels is more concrete, thus, easier to comprehend, and fewer hypernym levels indicate more abstract language that is harder to understand.

The remainder of this paper is organized as follows: Section 2 reviews the common indices that form the base of many traditional readability formulas and the criticism they have received. In Section 3, we review Prototype Theory and discuss how it can aid us in finding the lexical difficulty of a text. Section 4 is about methodology – how to identify basic level words and how to assess the validity of our method against other readability formulas. Section 5 reports the results of the assessment and discusses the strength and weaknesses of our approach. In this section, we also suggest what can be done in subsequent research.

2. Literature Review

In this section we first summarize the indices of traditional readability formulas, give an account of the criticism these formulas meet, and introduce the purpose of our study. Among the multitude of factors underlying the reading process, we will focus on the lexical index.

2.1 Indices of Readability

2.1.1 Vocabulary Difficulty

The earliest work on readability measurement goes back to Thorndike (1921) where word frequency in a corpus is considered an important index in computing vocabulary complexity. This is based on the logic that the higher the frequency of a word, the more common and easier it is. Followers of this logic compiled word lists that include often-used and seldom-used words where the presence or absence of particular words on the lists assesses vocabulary difficulty, thus text difficulty.

Vocabulary difficulty is also measured by word length in many formulas, *e.g.*, the Flesch formula (Flesch, 1943, 1948, 1950) and FOG formula (McCallum & Peterson, 1982), or in terms of number of syllables (Fry, 1968). This is based on another intuitive assumption that the longer a word is, the more difficult it is to comprehend (Bailin & Grafstein, 2001).

2.1.2 Syntactic Difficulty

Syntactic complexity is another index in many readability formulas (Chall & Dale, 1995). For Dale & Chall (1948), Flesch (1948), and McCallum & Peterson (1982), syntactic complexity boils down to the average length of sentences in a text, although they vary in how they determine and utilize sentence length. The formula designed by Heilman, Collins-Thompson, Callan, & Eskenazi (2007) is a more recent example of this type. They propose that grammar-based predictions can be combined with vocabulary-based predictions to produce more accurate predictions of readability for both first and second language texts. They also suggest that language technologies must account for morphological features in languages which have a rich morphology, an issue relevant to grammatical features.

Also taking account of syntactic complexity, Miltsakaki & Troutt (2007) bases their algorithm on three readability formulas: Lix, Rix, and Coleman-Liau. The number of sentences, words, long words (seven or more characters), and letters in the text are taken into account. Another example is Das & Roychoudhury's work (2006), which built a readability index for Bangla using average sentence length (total words/ total sentences) and number of syllables per 100 words (total syllables/ total words*100).

2.1.3 Semantic Difficulty

Semantic factors such as counting abstract words (Flesch, 1943; Cohen, 1975) and propositional density and inferences (Kintsch, 1974) have also been put into regression analyses of readability assessment. In addition to these projects, Wiener *et al.* (1990) proposes a scale based on ten categories of semantic relations, *e.g.,* temporal ordering and causality, for assessing the utterance complexity. The reliability of the semantic scale was confirmed when it was applied to compare the utterances of fourth-, sixth-, and eighth-grade children, where significant differences in semantic density were found on their scale.

Since 1920, more than fifty readability formulas have been proposed in the hopes of providing tools to measure readability more accurately and efficaciously (Crossley *et al*., 2007). Nonetheless, it is not surprising to see criticism over these formulas, given that reading is an extremely complex process.

2.2 Criticism of the Traditional Readability Formulas

Although classic readability formulas provide a quick and easy method of predicting readability, they are often criticized for being superficial, unstable, or unable to offer information about deeper levels of text processing (McNamara *et al*., 1996).

2.2.1 Criticism of Lexical Difficulty Measurement

Bailin & Grafstein (2001) question the validity of measuring vocabulary difficulty by the number of syllables per word or by the presence of words in a word list. They question the legitimacy of assessing vocabulary difficulty in terms of word length by showing that many mono- or bi-syllabic words are actually more esoteric, *i.e.* more unfamiliar, than longer polysyllabic terms. They also argue that the proposed link between readability and a vocabulary list of word frequency is narrowly based on the prerequisite that words in a language remain stable. The prerequisite, however, seems implausible as different socio-cultural groups have different core vocabularies and rapid cultural change makes many words out of fashion.

2.2.2 Criticism of Syntactic Difficulty Measurement

Bailin & Grafstein (2001) also point out the flaw of a simple equation between syntactic complexity and sentence length by giving the sample sentences as follows:

(1) I couldn’t answer your e-mail. There was a power outage.

(2) I couldn’t answer your e-mail because there was a power outage.

In terms of both absolute length and number of words, (2) is longer than (1), thus computed as more difficult by traditional readability formulas. Nevertheless, the subordinator

"because" in (2), which explicitly links the author's inability to e-mail to the power outage, actually aids comprehension. As such, the authors suggest that language-oriented criteria be proposed, including deviations from prescriptive grammar, style (relative clauses, garden-path phrases, left-branching structures, *etc.*), and required background knowledge.

2.2.3 Criticism of Statistical Legitimacy

The correlation between the indices and the measured variables was also challenged from the viewpoint of statistical legitimacy. Hua & Wang (2007) point out the methodological issue in the creation of the traditional readability formulas. The typical initial step is to select, as the criterion passages, standard graded texts whose readability has already been agreed upon. The next step is to sort out the factors that may affect the readability of the text. The factors that are highly correlated with the text difficulty are chosen as independent variables in regression analysis for forming a readability formula. The researchers, however, did not ascertain whether the factors incorporated into their regression model actually have a cause-effect relationship with the dependent variable, *i.e.*, readability. Word length, used to equate semantic complexity, and sentence length, used for syntactic complexity, are intuitively correlated with readability, but non-scientifically correlated. Therefore, the authors suggest that researchers first analyze the independent variables qualitatively to confirm their cause-effect relationship with readability.

Challenge also goes to the selection of criterion passages. Schriver (2000) suggests that readability formulas are inherently unreliable because they depend on criterion passages too short to reflect cohesiveness, too varied to support between-formula comparisons, and too text-oriented to account for the effects of lists, enumerated sequences, and tables on text comprehension.

2.3 Purpose of Research

The criticisms of the traditional readability formulas by the various authors have a lot in common. They all urge adoption of language-oriented criteria based on independent evidence and a closer re-examination of the genuine relationship between the variables and the texts. It is our belief that this can only be done if we take account of the deeper levels of text processing. Reading is a multidimensional process; our pilot study aims to examine how a reader interacts with a text at the **lexical** level. We propose that the hierarchical status of a lexical item in our mental lexicon is a possible factor that affects lexical comprehensibility. We further suggest that there is a basic level in the lexical hierarchy which is the easiest to comprehend and serves as a meaningful indicator of text readability. To that end, we resort to Prototype Theory, which was proposed and developed by Rosch *et al*. (1976), among others.

3. Prototype Theory and Lexical Difficulties

3.1 Prototype Theory

Prototype Theory was brought to cognitive linguistics by Rosch *et al.* (1976). The notion of *prototype* can be understood in two ways. First, prototype is used either to refer to object members that first come to one's mind in an association experiment, or to those that can be recognized faster than other category members in a verification task. For example, when asked to give an example of "bird", "robin" is more frequently cited than "ostrich". Various researchers (Rosch, 1978; Lakoff, 1986; Brown, 1990; Tversky, 1990) use different names to label the prototypical member – "best example of a category", "salient examples", "clearest cases of category membership", or "central and typical members." The other way to define prototype is from a genuinely cognitive viewpoint. Prototype can be viewed as a mental representation, specifically as some kind of cognitive reference point (Rosch & Mervis, 1975; Coleman & Kay, 1981; Lakoff, 1986). Taking the two viewpoints together, we can view prototype as the central member or the cognitive reference point which other members of the category are anchored to. Through the anchoring process, cognitive categories are formed.

The members within a particular cognitive category are anchored to the prototype with different parameters – whether the members are perceived as *gestalt*, how many category-wide attributes are shared by the members, and how homogeneous or heterogeneous the members are. The representation of a bird, for instance, does not consist of a set of features that all birds have. A robin or a penguin as a category member of the bird is anchored to the most typical or ideal category member of the bird (which may not exist in real life). Since a robin shares more of the features characteristic of a prototypical bird than a penguin shares, it is usually viewed by subjects as a better example of a bird.

The same mental anchoring process can be applied to broader human categorization of those readily identifiable organisms and objects that surround us (Ungerer & Schmid, 1996: 60). As a result, entities within the cognitive category of DOG can be categorized as a "dog", a "terrier", a "Scottish terrier", a "mammal", or an "animal". These cognitive categories are connected with each other in a hierarchical pattern. In this example, if we look at their relationship from the "bottom" of the hierarchy, Scottish terriers are subordinate to terriers, and terriers are subordinate to dogs. If we look at them from the "top" of the hierarchy, animals are viewed as superordinate to mammals, and mammals as superordinate to dogs.

Turning to early interpretations of the *basic level* from the psychological viewpoints by Brown (1958) and Kay (1971), the basic level is where human beings perceive the most obvious difference between the organisms and objects of the world. Imagine an everyday conversation where a person says "Who moved that piano?" The naming of an object with "piano" will not strike us as noteworthy until the alternative "Who moved that stringed

instrument?" is brought to our attention. Both terms are truth-conditionally adequate, but only the former is commonly used. The superordinate word "stringed instrument" is not used because its meaning encompasses many basic level words, *i.e*. many kinds of musical instruments. In our example, using the word "stringed instrument" is too vague to represent the object: "Stringed instrument" does not, as "piano" does, denote the most obvious difference of the object from the other objects in the world. Likewise, using a subordinate level word, *e.g*. a "grand piano", on a similar occasion is unusual except when the differentiation between different types of pianos is required.

In ranking typicality of objects, the basic level is where the largest bundles of naturally correlated attributes are available for categorization (Rosch *et al.*, 1976; Ungerer & Schmid, 1996: 67). In addition, the basic level is where gestalt perception occurs to the greatest extent, and this is particularly easy for prototypical examples. An "apple" has reddish or greenish skin, white pulp, and a round shape, while it is hard to pinpoint the features of "fruit". For a layman, hardly any significant features can be added to "crab apple".

The underlying cognitive anchoring process of the psychological reality of the basic level and the prototype are very similar. In the same way as other peripheral members of a category are anchored to the prototypical member, other non-basic levels, namely *superordinate* and *subordinate levels*, are anchored to the basic level. Ungerer and Schmid (1996: 72) point out the two are actually a kind of symbiosis underpinned by two interdependent principles: First, prototype categories are most fully developed on the basic level. Second, basic levels only function as they do because they are structured as prototypical categories.

The first principle can be explained by our earlier discussion on the basic level. Recall that this level offers the largest amount of correlated attributes, and the attributes are accumulated in their most completed form in the prototype and expressed by the category name (*e.g*., "Robin", as the typical example of the category BIRD, accumulates most correlated attributes of medium size, feathers, flying and singing ability, *etc*. in a complete form.)

As for the second principle, maximization of the efficiency of basic level categories by prototypes can be used to explain it (Rosch, 1977, 1978). That is, prototypes maximize the discontinuities or the distinctiveness of the basic level categories as they induce not only the greatest number of attributes shared inside the category, but the greatest number of attributes not shared by members of other categories. A typical example of a bird like "robin" can, while a non-typical example of the bird like "penguin" cannot, be easily distinguished from the category of fish because the latter shares more attributes with fish.

Developmentally, basic level categories are acquired earlier by children than their superordinate and subordinate words. Conceptually, the basic level category represents the

concepts humans interact with most readily. Applying the hierarchical structure of conceptual categorization to lexical comprehensibility, we suggest that a concept at the basic level, hence, the word that denotes the concept, which we call a basic level word, is easier for the reader than its superordinate and subordinate words. If this is correct, then one text should be easier than another if it contains more basic level words. As the three-leveled hierarchy refers specifically to nouns in Prototype Theory, we confine our current study to the nominal category only. The best tool to study the relevant hierarchical relations in a broad framework with computational techniques is WordNet.

3.2 WordNet – A Hierarchically-Structured Lexical Database of English

WordNet is a large online electronic lexical database of English. The words are interlinked by means of conceptual-semantic and lexical relations. Its underlying design principle has much in common with the hierarchical structure proposed in Prototype Theory. In the vertical dimension, the hypernym/hyponym relations among the nouns can be interpreted as hierarchical relations between conceptual categories. For instance, the direct hypernym of "apple" in WordNet is "edible fruit". One of the direct hyponyms of "apple" is "crab apple". Note, however, that hypernyms and hyponyms are relativized notions in WordNet. Theoretically speaking, any word may have hypernyms and hyponyms. "Crab apple," being a hyponym of "apple," is also a hypernym in relation to "Siberian crab apple". An ontological tree may well exceed three levels. There are no labels or ready-to-be-used statistical information in WordNet that tell us which nouns fall into the basic level category. In the following sections we try to retrieve the basic level nouns as defined in Prototype Theory and apply the results in assessing text readability.

4. Methodology

Three experiments were conducted. In the first experiment, we utilized the nouns used in Rosch *et al*'s experiments in order to discover their quantitative properties. The second experiment followed up the first one and tried to pinpoint the criteria of determining the quality of the nouns being basic. In the third experiment, we computed and compared the basic level noun ratios and readability scores of graded readings. Our results indicate that basic level noun ratios provide a robust and meaningful index of text readability.

4.1 Experiment 1

4.1.1 Design of Experiment 1

In our initial experiment, we examined the eighteen basic level words identified by Rosch *et al*. (1976: 388), checking their word length, lexical complexity, and their direct hypernyms as

well as direct hyponyms in WordNet. We speculate that a basic level word has these features: (1) It is relatively short (containing fewer letters than its hypernyms/hyponyms on average); (2) It is morphologically simple[1]; (3) It has more direct hyponym synsets than direct hypernym synsets[2]. Notice that some entries in WordNet are made up of more than one word. We assume that an item composed of two or more words is NOT a basic level word. A lexical entry composed of two or more words is a compound. The first word of a compound noun may or may not be a noun, and there may or may not be spaces or hyphens between the component words of a compound. For words having more than one sense, we focused only on the sense occurring in Rosch *et al*'s experiment. As an example, the noun "table" has six senses (*i.e.* synsets) in WordNet, but only the information in the sense of "a piece of furniture" is computed. Table 1 summarizes the results of Experiment 1.

Table 1. Eighteen basic level words in comparison with their direct hypernyms and direct hyponyms on word length, number of synsets, and morphological complexity*

Target word	Index of the inquired synset	Basic level		Direct hypernym			Direct hyponym		
		Word length	Morph. Complexity	Average word length	Number of synsets	Morph. Complexity	Average word length	Number of synsets	Morph. Complexity
guitar	0	6	A	18	1	B	8.8	6	A, B
piano	0	5	A	19	3	B	9.6	3	A, B
drum	0	4	A	20	1	B	7.4	8	A, B
apple	0	5	A	8.3	2	A, B	10.6	3	A, B
peach	2	5	A	8.7	2	B	N/A	N/A	N/A
grape	0	5	A	11	1	B	11.8	3	A, B
hammer	1	6	A	8	1	B	9.7	7	A, B
saw	1	2	A	8	1	B	8.7	7	A, B

[1] It has been noticed by Ungerer & Schmid (1996: 98) that basic level words tend to be short and mono-morphemic words, and that the superordinate and subordinate words tend to be longer and morphologically complex words. Nevertheless, to our knowledge, no systematic studies have been done prior to ours on these quantitative features.

[2] Both hyponyms and hypernyms are grouped into synsets in WordNet. A synset is a set of synonyms. The direct hyponyms of "guitar", for instance, are grouped into six synsets: (1) acoustic guitar, (2) bass guitar, (3) cittern, cithern, cither, citole, gittern, (4) electric guitar, (5) Hawaiian guitar, steel guitar, and (6) uke, ukulele. In contrast, "guitar" has only one direct hypernym synset, *i.e.* "stringed instrument". Accordingly, the number of direct hypernym synsets of "guitar" is 1, and the number of its direct hyponym synsets is 6, as shown in Table 1.

screwdriver	0	11	B	8	1	B	19.8	3	B
pant	1	4	A	7	1	A	8.9	18	A, B
sock	0	4	A	5.5	1	A	7.8	5	A, B
shirt	0	5	A	7	1	A	8.1	9	A, B
table	1	5	A	14.3	1	A	10.4	26	A, B
lamp	1	4	A	14.3	1	B	9.7	3	A, B
chair	0	5	A	4	1	A	10.7	15	A, B
car	0	3	A	14.5	1	A, B	8.3	31	B
bus	0	3	A	15	1	A, B	10.8	3	B
truck	0	5	A	14.5	1	A, B	8.7	11	B

*A refers to “single word”. B refers to “compound”.

4.1.2 Results of Experiment 1

The results confirm our prediction. First, the average word length (number of letters) of both the hypernyms and the hyponyms is much longer than that of the basic level words. Although many researchers have pointed out that absolute word length is not a meaningful indicator of lexical complexity (Davidson & Kantor, 1982; Bailin & Grafstein, 2001; Hua & Wang, 2007), to our knowledge no researchers so far have been able to propose a better algorithm to account for word length or be able to refute the intuition that word length plays a certain role in reflecting lexical difficulty.

Our results indicate that word length should be viewed in a relative sense; namely, there is a level in the lexical hierarchy that has the shortest word length in comparison with its higher and lower levels on average. That absolute word length is not a good index is manifested by the fact that in Table 1 some of the direct hypernyms are shorter than six letters in their average length, while some basic level words have only six letters. We have, however, observed consistency in the word length difference between levels of words within a hierarchy: The basic level words always contain the fewest number of letters[3]. The tendency is particularly strong when we compare the length of basic level words with the average length of their direct hyponyms[4].

[3] The tendency has only one exception, *i.e*. the direct hypernym of “chair” is shorter than “chair”.

[4] The word “screwdriver” seems to be an exception to the pattern we describe, as the average length of its direct hypernyms is shorter than the target word itself (8 < 11). Nevertheless, since “screwdriver” is a compound, *i.e*. composed of “screw” and “driver”, it is actually excluded by our basic level word criteria.

Our second finding from Experiment 1 is that these basic level words have many more direct hyponym synsets than direct hypernym synsets. Finally, in contrast to the basic level words which are morphologically simple, the direct hypernyms and the direct hyponyms are more complex. Many of the hypernyms are compounds. The hyponyms are even more complex. Every basic level word (except for "peach") has compounded hyponyms.

4.2 Experiment 2

4.2.1 Design of Experiment 2

Our first findings brought our attention to the relative length difference of the words at different levels and the disparity of their morphological structure. In particular, we found that basic level words display sharper contrast with their hyponyms than with their hypernyms, both in terms of word length difference and morphological structure complexity.

In this experiment, we set out to compute the difference between the length of the basic level words in Experiment 1 and the average length of their full hyponyms. We also examined the distribution of the compounds formed by the three levels of words - basic level words, their hypernyms, as well as their hyponyms. Additionally, we randomly came up with seven more words that appear to fall into the basic level category defined by Rosch *et al*. (1976). The results of this experiment are shown in Table 2.

Table 2. The twenty-five basic level words – Word length differences, compound ratios, and distribution of compounds

Hypernym / Basic level word / Hyponym	Index of the inquired synset	Average hyponym l. – target word l.	# of compounds / # of full hyponyms	Cpd ratio (%)	Number of compounds at hyponymous levels					
					1st level	*2nd level*	*3rd level*	*4th level*	*5th level*	*6th level*
stringed instrument	0	- 10.2	1 / 85	1	1	0	0	0		
guitar	0	**2.8**	5 / 12	**42**	5					
acoustic guitar	0	N/A	N/A	N/A						
keyboard instrument	0	- 9.0	0 / 35	0	0	0	0			
piano	0	**6.0**	8 / 16	**50**	4	4				
grand piano	0	1.9	3 / 8	38	3					
baby grand piano	0	N/A	N/A	N/A						

percussion instrument	0	- 12.5	0 / 68	0	0	0	0			
drum	0	**3.4**	5 / 14	**36**	5					
bass drum	0	N/A	N/A	N/A						
edible fruit	0	- 3.1	0 / 258	0	0	0	0	0		
apple	0	**5.5**	5 / 29	**17**	5	0	0			
crab apple	0	3.6	2 / 8	25	2					
Siberian crab	0	N/A	N/A	N/A						
edible fruit	0	- 3.1	0 / 258	0	0	0	0	0		
peach	2	**N/A**	N/A	**N/A**						
N/A	N/A	N/A	N/A	N/A						
edible fruit	0	- 3.1	0 / 258	0	0	0	0	0		
grape	0	**4.5**	6 / 17	**35**	3	2	1			
muscadine	0	N/A	N/A	N/A						
hand tool	0	1.0	0 / 217	0	0	0	0	0		
hammer	1	**3.9**	7 / 16	**44**	7	0				
ball-peen hammer	0	N/A	N/A	N/A						
hand tool	0	1.0	0 / 217	0	0	0	0	0	0	
saw	1	**5.7**	25 / 30	**83**	13	12	0			
bill	7	N/A	N/A	N/A						
hand tool	0	1.0	0 / 217	0	0	0	0	0	0	
screwdriver	0	**8.8**	4 / 4	**100**	4					
flat tip screwdriver	0	N/A	N/A	N/A						
garment	0	1.0	4 / 306	1	3	1	0	0	0	
pant	1	**4.9**	10 / 49	**20**	9	1				
bellbottom trousers	0	N/A	N/A	N/A						
hosiery	0	1.7	0 / 28	0	0	0				
sock	0	**3.8**	5 / 13	**38**	5					
anklet	0	N/A	N/A	N/A						
garment	0	1.0	4 / 306	1	3	1	0	0	0	
shirt	0	**3.2**	8 / 17	**47**	8	0				

camise	0	N/A	N/A	N/A						
furniture	0	0.0	4 / 244	2	4	0	0	0	0	
table	1	**5.1**	36 / 77	**47**	29	7	0	0		
altar	0	N/A	N/A	N/A						
source of illumination	0	- 11.9	0 / 107	0	0	0	0	0	0	
lamp	1	**6.0**	4 / 4	**100**	3	1				
Aladdin"s lamp	0	N/A	N/A	N/A						
seat	2	5.0	7 / 101	6.9	3	3	1	0		
chair	0	**5.6**	31 / 48	**65**	17	14	0			
armchair	0	2.7	0 / 10	0	0	0				
captain's chair	0	N/A	N/A	N/A						
motor vehicle	0	- 4.3	0 / 151	0	0	0	0	0		
car	0	**5.3**	21 / 75	**28**	19	2				
amphibian	0	1.0	0 / 2	0	0					
public transport	0	- 6.7	0 / 38	0	0	0	0			
bus	0	**7.8**	3 / 5	**60**	3					
minibus	0	N/A	N/A	N/A						
motor vehicle	0	- 4.3	0 / 151	0	0	0	0	0		
truck	0	**4.5**	15 / 48	**31**	10	5	0			
dump truck	0	N/A	N/A	N/A						
canine	1	4.3	0 / 287	0	0	0	0	0	0	0
dog	0	**8.0**	50 / 235	**21**	11	21	16	2	0	
puppy	0	N/A	N/A	N/A						
feline	0	3.1	0 / 123	0	0	0	0			
cat	0	**6.1**	35 / 87	**40**	4	30	1			
domestic cat	0	- 4.0	0 / 32	0	0					
kitty	2	N/A	N/A	N/A						
publication	0	- 1.3	2 / 192	1	0	1	0	0	0	
book	0	**6.5**	38 / 139	**27**	17	11	7	3	0	
authority	6	- 1.0	0 / 1	0	0					
last word	0	N/A	N/A	N/A						

language unit	0	- 3.0	0 / 290	0	0	0	0	0	0	0
word	0	**6.5**	35 / 185	**19**	28	7	0	0	0	
anagram	6	1.0	0 / 1	0	0					
antigram	0	N/A	N/A	N/A						
material	0	1.1	17 / 591	2.9	15	2	0	0		
paper	0	**4.1**	59 / 173	**34**	40	18	1			
card	0	3.1	14 / 57	25	6	8				
playing card	0	- 5.4	0 / 49	0						
movable barrier	0	- 6.2	0 / 44	0	0	0	0			
door	0	**5.8**	18 / 23	**78**	13	5				
car door	0	3.0	0 / 2	0	0					
hatchback	0	N/A	N/A	N/A						
leaf	1	4.7	2 / 21	10	2	0	0			
page	0	**5.0**	5 / 18	**28**	5	0				
full page	0	N/A	N/A	N/A						

In the first column, the basic level words (*e.g.* "guitar") are boldfaced, with the (or one of the) direct hypernym(s) (*e.g.* "stringed instrument") given above and its first-occurring direct hyponym (*e.g.* "acoustic guitar") placed under it. When the basic level word has more than one level of hyponym, the first word occurring at the second hyponymous level was also examined such that the word "hatchback" was under "car door".

As in Experiment 1, with respect to words having more than one sense, we focused only on the sense defined in Rosch *et al*. (1976). As an example, the noun "table" has six senses (*i.e*. synsets) in WordNet, but only the information in the sense of "a piece of furniture" is computed. Which synset conforms to the sense in Rosch *et al*. (1976) was decided manually. In WordNet, each sense of a word is indexed numerically. We put the index of the inquired synset in the second column. Notice that the first sense of a word has the number 0, and the second sense, the number 1, and so on and so forth.

All other information was retrieved by a program we wrote based on NLTK-0.9.5, which was downloaded at http://www.nltk.org/. (NLTK had been updated to version 0.9.9 by the time of revision of this paper.) Our own program can be downloaded at http://lope.eng.ntnu.edu.tw/lopedia/index.php/Image:Compound_ratios_and_word_length_difference_in_WordNet.doc#filelinks). We set the hyponym depth in our program at 100 levels.

The third column "Full hyponym length minus target word length" computes the difference between the length of the target word and the average length of its full hyponyms.

The word "stringed instrument" has, for example, 85 hyponyms in total with their average length being 8.79 letters. The length difference is 8.79 minus 19, which equals -10.2. A negative value in this column thus means that the target word is longer than the average length of its full hyponyms. On the other hand, a positive value conveys longer average hyponym length than the target word.

The fourth column computes the ratios of compounds composed of the target word in the full hyponyms. Our program searches the full hyponyms for compounds that are formed by the target words[5]. Such a compound may end with the target word, which constitutes the major compounding pattern we have observed, *e.g*. "school bus" is a compound hyponym of "bus". The other way to form a compound hyponym is to start with the target word. The only examples we know, however, are "icefall," "ice pack," and "ice shelf" in the second synset of "ice". In light of the existence of compounds like those, we also include this compounding template in our program. The program virtually searches for compounds that contain the target word, assuming that the target word may occur in the front, middle, or end position of the compound.

As an example of the compounding behavior and the computation of compound ratio, among the twelve (full) hyponyms of "guitar," five are compounds formed by "guitar". They are "acoustic guitar," "bass guitar," "electric guitar," "Hawaiian guitar," and "steel guitar". The compound ratio of "guitar" is accordingly 5/12 = 42%. By contrast, only one hyponym of the full eighty-five hyponyms of "stringed instrument" is a compound containing "stringed instrument" (*i.e*. "bowed stringed instrument"), and its compound ratio is 1/85 = 1%. As for "acoustic guitar," it has no hyponyms. These compound ratios are given in the fifth column.

We also keep record of the levels where compounds occur, which we display in the rightmost columns.

4.2.2 Results of Experiment 2

Several regular patterns can be observed in Table 2. In terms of word length difference, the basic level words show the greatest positive values across the board. Each basic level word enjoys greater length difference than its direct hypernym and its first-occurring hyponym do. On the other end of the length difference spectrum are the direct hypernyms. Of the twenty-five direct hypernyms, ten have negative values. Only four have positive length differences greater than 3. This is likely to result from the fact that the direct hypernyms in Table 2 are mostly compound words, thus, tend to be long. If we go one level higher, the length difference may decrease. For example, the direct hypernym of "edible fruit" is "fruit",

[5] When a word is compounded with another word to form a compound, this compounded word becomes a hyponym of the target word and is unlikely to become a hypernym of the target word.

whose length is much shorter.

The most significant finding from Experiment 2 is that basic level words have the highest compound ratios. In comparison with their hypernyms and hyponyms, they are much more frequently used to form compounds. Although some hyponyms like “grand piano” and “crab apple” also have high compound ratios, they should not be taken as basic level items because these compounds often contain the basic level words themselves (*e.g.* “Southern crab apple” contains “apple”), indicating that the ability to form compounds is actually inherited from the basic level words.

Our data pose a challenge to Prototype Theory in that a subordinate word of a basic level word may act as a basic level word itself. The word “card,” a hyponym of “paper,” is of this type. With its high compound ratio of 25%, “card” may also be deemed to be a basic level word. This fact raises another question as to whether a superordinate word may act as a basic level word as well. One might, however, object that it is doubtful whether “card” is really subordinate to “paper” in the framework of Prototype Theory. That is to say, it takes independent evidence to prove that the hyponyms of these twenty-five basic level words in WordNet correspond to the subordinate words defined by Prototype Theory and that the hypernyms correspond to the superordinate words. We leave this issue aside for reasons that will be clear when we describe the design of the next experiment. At this moment, suffice it to say that the way we identify basic level words in WordNet is not based on how many levels of hyponyms or hypernyms a word has or on which specific level in the hierarchy a word falls.

Many of the basic level words in Table 2 have three or more levels of hyponyms. This indicates that what is cognitively basic may not be low in the ontological tree. A closer look at the distribution of the compounds across the hyponymous levels reveals another interesting pattern. Basic level words have the ability to permeate two to three levels of hyponyms in forming compounds. In contrast, words at their hypernymous levels do not have such ability, and their compounds mostly occur at their direct hyponymous levels only. Words at their hyponymous levels rarely, if ever, form compounds.

4.3 Experiment 3

4.3.1 Design of Experiment 3

The goal of this experiment is to show that whether a word belongs to the basic level affects its comprehensibility, which in turn affects the readability of a text. If this is correct, when all other factors are equal, an easy text should contain more basic level words than a difficult text. Put in fractional terms, we attempt to show that the proportion of basic level words in a text is correlated with the readability of the text.

To achieve this goal, we need independent readability samples to compare with our prediction. Nevertheless, as readability is a subjective judgment that may vary from one person to another, such independent samples are extremely difficult, if possible, to obtain. In this study, we resorted to a pragmatic practice by selecting the online graded readings for American children and texts in English textbooks for senior high school students in Taiwan. Five open source readings ranging from grade one to twelve from edHelper.com (http://www.edhelper.com/ReadingComprehension.htm) were randomly selected for Experiment 3. Three textbooks from Sanmin Publishing Co., each used in the first semester of a different school year, were also used for this experiment. We tried to choose the same type of text, so that text type would not act as noise. All three high school English texts are informational. Due to the great divergence between the levels of children's readings, however, it was not easy to be strict with text types. Furthermore, since we do not have the facilities to run large-scale experiments yet, we limited the scope to around four-hundred-word texts at the Taiwanese high school level, and approximately two-hundred-and-fifty word texts at the American children's level. All selected readings are appended in Appendix A and B. Using the same program as in Experiment 2, we searched WordNet 3.0 for all the nouns occurring in these texts, except for proper names and pronouns. We referred only to the words in the particular sense occurring in the selected readings. We know that this practice, if used in a large-scale study, is applicable only if sense tagging is available.

Based on the results of the two preliminary experiments, we argue that the basic level noun index includes at least the following two quantitative features: (1) A basic level noun has great ability to form compounded hyponyms; (2) The length of a basic level noun is shorter than the average word length of its full hyponyms. These characteristics can be further simplified as the **Filter Condition** to pick out basic level nouns:

(1) Compound ratio (*i.e.* the number of the hyponyms which contain the target word divided by the number of the target word's full hyponyms) $\geqq$ 20%;

(2) Length difference (*i.e.* the average length of the target word's full hyponyms minus the length of the target word) $\geqq$ 2.

We set the compound ratio threshold at twenty percent for the following reasons. On the one hand, the compound ratios of all the basic level words in Experiment 2, except for "apple," "pant," and "word," are higher than twenty-five percent. On the other hand, these basic level words are derived from the psycholinguistic experiments by Rosch *et al.* (1976) which were designed to markedly manifest the human conceptual structure of categorization. Due to the special purpose of their experiments, these words are supposed to be the most typical basic level words in the English vocabulary. The quantitative data obtained using these words should be fine-tuned to a lower level to capture the representativeness of the other not-so-typical basic level words. As the compound ratios of "apple," "pant," and "word" are

all near twenty percent, we approximate the threshold of the compound ratio to be twenty percent. In our future research, with more training data, the threshold will be further weighted.

The reason for setting up the condition of word length difference at two letters is the same as for the setting of the compound ratio: The basic level words in Experiment 2 are the most typical, therefore, on the upper end of the spectrum. Even though nineteen of the twenty-five words are shorter than their average hyponym by at least four letters, we set the word length difference condition at two letters. This threshold should also be further weighted in future research.

Note in passing that the second criterion differs fundamentally from the commonly used criterion of word length. Ours compares the target word with the average length of its full hyponyms. In our study word length is measured in relative terms: The word length difference is an index, not the word length itself.

Based on the two criteria of our filter condition, the information for each noun we need include the length of the target word, the average word length of its full hyponyms, the number of its full hyponyms, and the number of compounds of the target word, *i.e*. how many hyponyms of the word are compounds formed by the word. All computed values for each noun in the selected readings can be found in Appendix C and D. Words that pass the filter condition are displayed in red color with their compound ratios and length differences being boldfaced.

The next step of Experiment 3 was to compute the basic level noun ratios of the selected readings. Basic level noun ratios were obtained by dividing the number of basic level noun types in a text by the number of all noun types of the text. For example, the easiest text in our randomly selected online readings for American children ("Wash Your Hands" in Appendix A) contains 21 noun types (excluding proper names and pronouns), and 12 of them (*i.e*. the red items in the first table of Appendix C) reach the basic level noun threshold. The basic level noun ratio of this reading is accordingly 12/21 = 57.1%. Using the online software *Readability Calculations* (http://www. micropowerandlight.com/rd.html), we also obtained the scores of these texts computed by several readability formulas. These scores were then compared with the basic level noun ratios. We report and discuss the results of Experiment 3 in the next section.

5. Results and Discussion

Table 3 shows the raw scores and z-scores of the American children's readings and the Taiwanese high school texts calculated in terms of basic level noun ratios and by several readability formulas. With respect to the children's texts, Level 1 is the easiest level and Level 12 is the hardest.

Table 3. Raw- and z-scores of children and high school English texts computed by basic level noun ratios and several readability formulas

Measurement	Score	**Level 1 ~ 2**	**Level 3 ~ 4**	**Level 4 ~ 6**	**Level 7 ~ 8**	**Level 9 ~ 12**	**Book 1 Lesson 2**	**Book 3 Lesson 1**	**Book 5 Lesson 1**
Basic Level Word Ratio (%)	Raw score	57.1	48.3	32.6	28.6	21.1	39.3	26.1	25.0
	Z-score	*1.78*	*1.08*	*-0.17*	*-0.49*	*-1.09*	*0.36*	*-0.69*	*-0.78*
Dale_Chall	Raw score	4.7	6.3	7.2	6.6	7.8	4.5	7.0	7.1
	Z-score	*-1.42*	*-0.08*	*0.67*	*0.17*	*1.17*	*-1.59*	*0.5*	*0.59*
Flesch Grade Level	Raw score	0.3	4	6	6.3	8.1	1.5	7.9	9.3
	Z-score	*-1.59*	*-0.44*	*0.18*	*0.27*	*0.83*	*-1.21*	*0.77*	*1.2*
FOG	Raw score	2.8	9.7	19.1	13.2	18	*-1.1*	*17*	*21.4*
	Z-score	*-1.59*	*-0.56*	*0.85*	*-0.03*	*0.69*	*-1.1*	*0.54*	*1.2*
Powers	Raw score	3.3	4.7	5.5	5.2	6	*3.8*	*6.0*	*6.3*
	Z-score	*-1.65*	*-0.37*	*0.37*	*0.09*	*0.83*	*-1.19*	*0.83*	*1.1*
SMOG	Raw score	3.9	7.5	9.8	9.1	10.1	*9.8*	*11.9*	*11.9*
	Z-score	*-2.06*	*-0.67*	*0.21*	*-0.06*	*0.33*	*0.21*	*1.02*	*1.02*
FORCAST	Raw score	6.4	9	9.4	9.1	11.2	*7.4*	*11.5*	*10.9*
	Z-score	*-1.63*	*-0.2*	*0.02*	*-0.14*	*1.01*	*-1.08*	*1.18*	*0.85*
Spache	Raw score	1.8	3.0	3.6	3.8	4.2	*2.2*	*3.6*	*4.5*
	Z-score	*-1.63*	*-0.36*	*0.28*	*0.49*	*0.92*	*-1.21*	*0.28*	*1.23*

Diagrammatically, it is clear in Figure 1 that the basic level noun ratios of the American children's texts decrease inversely proportionally to the difficulty levels of the selected readings. Readability scores of the same texts calculated by the traditional formulas contain more ups and downs.

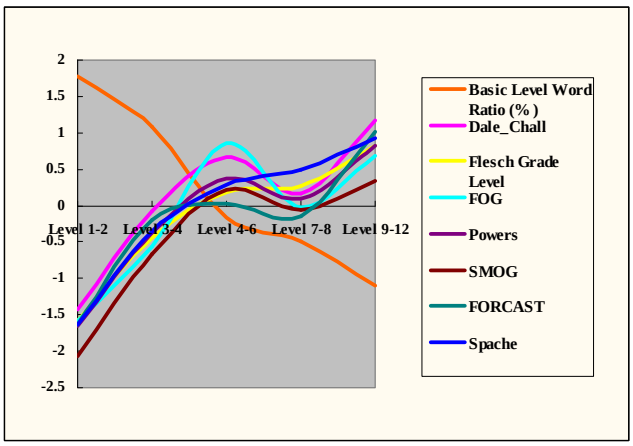

Figure 1. Readability of children's texts computed by basic level noun ratios and several readability formulas

As for the Taiwanese high school English texts, Figure 2 shows that Book 3, Lesson 1 and Book 5, Lesson 1 are rated similarly both in terms of basic level word ratios and by most of the readability formulas. We suspect that the textbooks are not well differentiated according to the levels of the students.

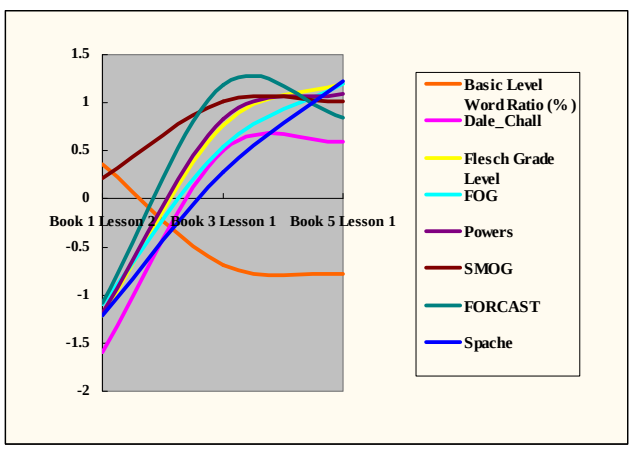

Figure 2. Readability of high school English texts computed by basic level noun ratios and several readability formulas

Overall, the basic level word ratios obtained in our study, both for high school students in Taiwan and for American children, conform to the levels of these texts, proving the usefulness of the basic level word concept in the assessment of readability.

The hierarchical relations in WordNet have also been utilized in Cohmetrix (McNamara *et al.* version 2.0), an online readability assessment software. It also uses the hierarchical

relations as an index of conceptual difficulty:

> "A word having more hypernym levels is more concrete. A word with fewer hypernym levels is more abstract."

What the Cohmetrix calculates is the mean levels above the words (nouns, verbs, and adjectives). A word at the bottom of an ontological tree in WordNet is deemed in Cohmetrix to be the most concrete. Since what is more concrete is generally believed to be simpler, a word at the lowest level is viewed as the easiest conceptually. The higher the level reaches, the more abstract, hence, conceptually more challenging for the human processor it becomes. This may seem intuitively sound, but our study has clearly shown that the relations between lexical items in a hierarchy are not like a ladder, a metaphor that captures what the Cohmetrix calculation seems to imply of the relations between the lexical items in our mental lexicon. We used the online software of Cohmetrix and obtained the scores in Table 4.

Table 4. Raw- and z-scores of lexical conceptual difficulty computed by Cohmetrix

Measurement	Score	**Level 1 ~ 2**	**Level 3 ~ 4**	**Level 4 ~ 6**	**Level 7 ~ 8**	**Level 9 ~ 12**	**Book 1 Lesson 2**	**Book 3 Lesson 1**	**Book 5 Lesson 1**
Mean hypernym values of nouns	Raw score	5.721	5.637	4.831	5.55	4.623	4.935	4.65	4.823
	Z-score	*1.36*	*1.17*	*-0.58*	*0.99*	*-1.03*	*-0.35*	*-0.97*	*-0.59*

Table 4 shows that the mean hypernym values of the nouns in these texts are not correlated with the text levels. This is illustrated by the sharp up and down in Figure 3 and the big curve in Figure 4.

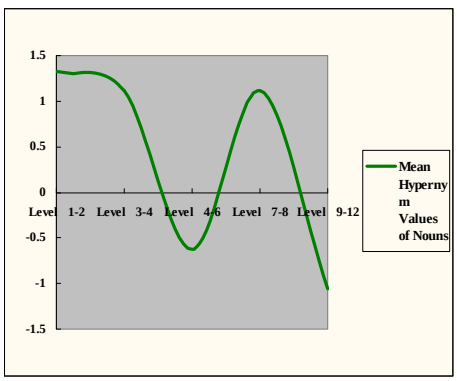

Figure 3. Readability of children's texts computed in terms of mean hypernym values of words

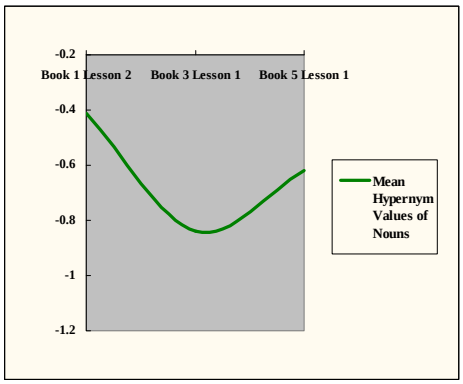

Figure 4. Readability of high school English texts computed in terms of mean hypernym values of words

These data suggest that a calculation that takes the lowest lexical hierarchical level as the most basic apparently misses the conceptual divide in our mental lexicon marked up by the basic level.

This paper is just the first step in assessing readability by lexical relations retrieved from WordNet based on conceptual categorization theories. Our study of readability assessment takes an approach that deviates remarkably from the traditional readability formulas. It looked for language-oriented variables that truly correlate with the reading process based on independent evidence. Prototype Theory, which is rooted in cognitive psychology and linguistics, gives us the idea that words form levels in our mental lexicon, with each level having its own characteristics and, accordingly, varied comprehensibility. It is our belief that these qualitative properties proposed in linguistic theories have corresponding quantitative features and retraceable distributions. The aim of our research is to find out these features by means of computational linguistic approaches and apply them in the assessment of text readability. The electronic lexical database of WordNet is an excellent tool to test our hypothesis. The results of our experiments are stimulating, but at the same time pose more challenges than achievements.

The filter condition of basic level nouns proposed in this study still leaves room to be fine-tuned and improved at least in two respects. First, the two criteria of compound ratios and word length difference have been used as sufficient conditions. More experiments will be designed for weighting these parameters in our future research. Specifically, we will study the distribution of compound words and the distribution of hyponyms over English words as one of our reviewers has pointed out that the distributions of these basic quantities affect the ratios. Furthermore, as another reviewer points out, the parameters must be able to be transformed into a scale in the future. Second, in addition to the lexical relations proposed in this study, there are presumably other relations between basic level words and their

hypernyms/hyponyms that are retrievable via WordNet and other databases. These relations, if found, can further modify the basic level word criteria proposed in this study.

Doubts can be raised as to whether all basic level words are equally readable or easy. Can it be that some basic level words are in fact more difficult than others and some hypernyms/hyponyms of certain basic level words are actually easier than certain basic level words? Are basic level words frequent words in general? Can we substitute frequency for the quality of being basic if the two criteria have approximately the same indexing power? This question can be extended to whether the hierarchical relations between the lexical units in WordNet are correlated with word frequency, and if so, in what ways. We will try to answer these questions in a study of larger scale.

The examined words in this study are all nouns. Can we find relations between verbs, adjectives, and even adverbs like the hypernym/hyponym relations within the various levels of nouns? The tentative answer to this question is yes and no. Take the example of the verb "run". It has hypernyms in WordNet ("speed," "travel rapidly," *etc*.). It also has subordinate lexical relations called "troponym," which are similar to hyponyms of nouns. English verbs, admittedly, do not constitute compounds as often as English nouns, but other lexical relations may exist between the verbs, and the relations are likely to be retrievable.

As Bailin & Grafstein (2001) suggest, lexical difficulty assessment should take into account the socio-cultural groups whose core vocabulary and background knowledge differ considerably in specific fields. Our initial speculation in this respect is that every academic and professional discipline has its own set of basic level words. These words may be highly infrequent in everyday use of the language, but form the fundamental layer in the jargon in its own sphere. A truly useful readability measurement tool thus should correspond to the text category and meet the readers' personal needs.

Future readability assessment tools should also be able to report not only the difficulty levels of the texts according to the readers' background knowledge but also the difficulty itself. In the lexical dimension, the tool should highlight the high level vocabulary for the readers. An algorithm like our current application is working exactly in this direction.

Laying out the groundwork for further research, we aim to tackle the following issues as well. All traditional readability formulas implicitly suppose an isomorphic relation between form and meaning as if each word has the same meaning no matter where it occurs. We acknowledge that one of the biggest challenges of measuring readability is to disambiguate the various senses of a word in text as the same word may have highly divergent readability in different senses. Another tacit assumption made by the traditional readability formulas is that the units of all lexical items are single words. This assumption overlooks many compounds and fixed expressions, affecting the validity of these formulas. This raises the issue of

segmentation. It is clear that the rating process applied in this study cannot be fully automated without successful segmentation of the text.

Although the small scale size of our experiments makes the validity of the results challengeable, its findings have provided the outlook of a large-scale project in the future. It has opened up a new approach to the assessment of text readability.

References

Bailin, A. & Grafstein, A. (2001). The linguistic assumptions underlying readability formulae: A critique. *Language and Communication,* 21(3), 285-301.

Brown, R. (1958). How shall a thing be called? *Psychological Review*, 65(1), 14-21.

Chall, J., & Dale, E. (1995). *Readability Revisited: The New Dale-Chall Readability Formula.* Cambridge, Massachusetts: Brookline Books.

Cohen, J. H. (1975). The effects of content are material on cloze test performance. *Journal of Reading,* 19(3), 247-250.

Coh-Metrix (Version 2.0) (Software). Memphis, TN: University of Memphis, Institute for Intelligent Systems. Available from http://cohmetrix.memphis.edu/cohmetrixpr/index.html

Coleman, L., & Kay, P. (1981). Prototype semantics: The English word "lie". *Language,* 57(1), 26-44.

Crossley, S. A., Dufty, D. F., McCarthy, P. M., & McNamara, D. S. (2007). Toward a new readability: A mixed model approach. In *Proceedings of the 29th Annual Conference of the Cognitive Science Society,* Austin, TX: Cognitive Science Society, 197-202.

Crossley, S. A., Greenfield, J., & McNamara, D. S. (2008). Assessing text readability using cognitively based indices. *Tesol Quarterly,* 42(3), 475-493.

Dale, E., & Chall, J. (1948). Formula for predicting readability. *Educational Research Bulletin* 27, 37-53.

Das, S., & Roychoudhury R. (2006). Readability modelling and comparison of one and two parametric fit: A case study in Bangla. *Journal of Quantitative Linguistics,* 13, 17-34.

Davison, A., & Kantor, R. (1982). On the failure of readability formulas to define readable texts: A case study from adaptations. *Reading Research Quarterly*, 17, 187-209.

edHelper. com, Reading Comprehensions, http://edhelper.com/ReadingComprehension.htm.

Fellbaum, C. (1998). *WordNet: An Electronic Lexical Database*. Cambridge, MA: The MIT press.

Flesch, R. (1943). *Marks of Readable Writing.* Ph.D. thesis.

—— (1948). A new readability yardstick. *Journal of Applied Psychology*, 32, 221-233.

—— (1950). Measuring the level of abstraction. *Journal of Applied Psychology*, 34, 384-390.

Fry, E. (1968). A readability formula that saves time. *Journal of Reading*, 11(7), 265-71.

Kay, P. (1971). Taxonomy and semantic contrast. *Language*, 47, 866-887.

Kintsch, W. (1974). *The Representation of Meaning in Memory*. Hillsdale, NJ: Erlbaum.

Heilman, M., Collins-Thompson, K., Callan, J., & Eskenazi, M. (2007). Combining lexical and grammatical features to improve readability measures for first and second language texts. In *Proceedings of the HLT/NAACL Annual Conference*, Rochester U.S.A., 460-467.

Hua, N. & Wang, G. (2007). Lun chuantong keduxing gongshi de bu-kexuexing. [On the non-scientific aspects of traditional readability formulae], *KaoShiZhouKan*, 18, 119-120.

Just, M. A., & Carpenter, P. A. (1987). A theory of reading: From eye fixations to comprehension. *Psychological Review*, 87, 329-354.

Lakoff, G. (1986). Classifiers as a reflection of mind. In C. G. Craig (Ed.), *Noun Classes and Categorization*, 13-52. Amsterdam: John Benjamins.

McCallum, D. R., & Peterson, J. L. (1982). Computer-based readability indices. In *Proceedings of the ACM '82 Conference*.

McNamara, D. S., Kintsch, E., Butler-Songer, N., & Kintsch, W. (1996). Are good texts always better? Interactions of text coherence, background knowledge, and levels of understanding in learning from text. *Cognition and Instruction*, 14, 1-43.

McNamara, D. S., Louwerse, M. M., & Graesser, A. C. (2002). Coh-Metrix: Automated cohesion and coherence scores to predict readability and facilitate comprehension. Unpublished technical report: University of Memphis.

Miltsakaki, E., & Troutt, A. (2007). Read-X: Automatic evaluation of reading difficulty of web text. In T. Bastiaens & S. Carliner (Eds.), *Proceedings of World Conference on E-Learning in Corporate, Government, Healthcare, and Higher Education 2007*, Chesapeake, VA: AACE, 7280-7286.

NLTK version 0.9.5 (Natural Language Toolkit) http://www.nltk.org/Home.

Perfetti, C. A. (1985). *Reading Ability*. Oxford: Oxford University Press.

Readability Calculations, Software, http://www.micropowerandlight.com/rd.html, Micro Power & Light co, downloaded on December 5, 2008.

Rayner, K., & Pollatsek, A. (1994). *The Psychology of Reading.* Englewood Cliffs, NJ: Prentice Hall.

Rosch, E., & Mervis, C. (1975). Family resemblances: Studies in the internal structures of categories. *Cognitive Psychology*, 7, 573-605.

Rosch, E., Mervis, C., Gray, W., Johnson, D., & Boyes-Braem, P. (1976). Basic objects in natural categories. *Cognitive Psychology*, 8, 382-439.

Rosch, E. Human categorization. In N. Warren (Ed.), *Advances in Cross-cultural Psychology (Vol. 1)*. London: Academic Press.

—— (1978). Principles of categorization. In E. Rosch & B. B. Lloyd (Eds.), *Cognition and Categorization*, Social Science Research Council (U.S.).

Rubin, A. (1985). How useful are readability formulas? In J. Osborn, P. T. Wilson, & R. C. Anderson (Eds.), *Reading Education: Foundations for a Literate America*, Lexington, MA: Lexington Books, 61-77.

Schriver, K. A. (2000). Readability formulas in the new millennium: What's the use? *ACM Journal of Computer Documentation*, 24(3), 138-140.

Thorndike, E. L. (1921). *The Teacher's Word Book.* New York: Teacher's College Press.

Tversky, B. (1990). Where partonomies and taxonomies meet. In S. L. Tsohatzidis (Ed.), *Meanings and Prototypes: Studies on Linguistic Categorization*. London, 334-344.

Ungerer, F., & Schmid, H. J. (1996). *An Introduction to Cognitive Linguistics*. London/New York, Longman.

Wiener, M., Rubano, M., & Shilkret, R., (1990). A measure of semantic complexity among predications. *Journal of Psycholinguistic Research*, 19(2), 103-123.

WordNet, version 3.0. (2006). Princeton, N.J.: Princeton University. Retrieved from World Wide Web: http://wordnet. princeton.edu/perl/webwn?s=word-you-want.

Appendix A: Online graded readings for American children (Each of 240 words)

Title: Wash Your Hands / Reading level suggested by edHelper: Grades 1-2

Downloaded at http://edhelper.com/ReadingComprehension_29_156.html

It was winter. It was fun in the snow. It was fun playing games. It was not fun being sick. A lot of kids had been sick at school. Some had colds. Some had the flu. Some hurt in the belly. The nurse came to visit each class. She talked about germs. She wanted to stop the germs. She gave tips to keep the germs away. The nurse came to Gabby's class. "Hi!!" she said. "How are you today?" "Fine," the children said. "Have any of you been sick this winter?" the nurse asked. Most of them raised their hands. "Do any of you want to be sick again?" the nurse asked. The kids shook their heads. It is no fun to be sick. "What makes you sick?" the nurse asked. "Germs!" the kids yelled. "That's right!" the nurse said. "What is the best thing you can do to keep germs away?" the nurse asked. The kids did not know what to say. The nurse smiled. "This is an easy one," she said, but the kids were still not sure. "Wash your hands," the nurse told them. "Germs are all over. You spread germs when your hands are dirty. Maybe you touched someone who was sick. Maybe you sneezed or coughed. Maybe you touched a dirty diaper. Maybe some food you touched had germs. If you don't wash the germs away and you touch your eyes or nose or mouth, you may get sick."

Title: The MDA Carnival Package Arrives / Reading level suggested by edHelper: Grades 3–4

Downloaded at http://edhelper.com/ReadingComprehension_29_183.html

Eli ran through the back door. "It's here, it's here," he said, waving a large envelope in the air. "What's here?" asked his sister, Sarah, as she reached into the cabinet to grab a package of cookies. "The carnival kit," said Eli as he sat down at the kitchen counter and started tearing the package apart. Sarah poured herself a glass of milk and put some cookies on a plate. Then she walked over to the counter, placed her snack down, and sat in the chair next to Eli, watching him pull the contents out of the package. "Why do you need a carnival kit?" "I'm going to have a Muscular Dystrophy Association Carnival," said Eli proudly. "It's going to raise lots of money for the MDA, just like Jerry Lewis." "Well la-de-da," said Sarah snippily. "I doubt you'll raise millions of dollars!" Eli glared at his sister. "Maybe not millions, but I'll bet I can raise thousands!" Sarah swallowed a bite of her cookie and washed it down with a gulp of milk. "You have big dreams!" She shook her head. Then she grabbed another cookie and popped it into her mouth. As Sarah continued to eat her snack, Mom came in the back door with an armload of groceries. "Hi, Mom," said Eli. "Look what came today!" "The carnival package!" said Mom, placing the bags of groceries on the counter. "That's great. Have you looked it over yet?"

Title: Physical Therapists / Reading level suggested by edHelper: Grades 4–6

Downloaded at http://edhelper.com/ReadingComprehension_29_192.html

Marilyn was a great-grandmother. One day she took a terrible fall. She needed surgery for her broken hip. She went to a rehabilitation hospital to heal. Who helped her to walk again? A physical therapist did. Joseph was walking to his car. He slipped on a patch of ice. He broke his leg. He needed to use crutches for a bit. Who helped him learn to use them? A physical therapist did. Nicholas is a year old. He was born with a disorder that has hindered his motor development. He attends a special gym three days a week where he can strengthen his muscles by playing with balls, benches, swings, and slides. Who helps him play and grow strong? A physical therapist does. So what is physical therapy? It is a special medical treatment that helps individuals move their bodies. Who does it help? It is meant for those with disabilities, illnesses, or injuries that interfere with movement. It is meant to keep the effects of a disability to a minimum, to help someone feel better, and to speed up recovery. Physical therapy is prescribed by doctors in the orthopedic, neurological, heart, and respiratory fields. Physical therapy has been around for a long time. Ancient people believed in parts of it. It became popular in the United States after the outbreak of World War I. The first school of physical therapy was at the Walter Reed Army Hospital in Washington,

Title: Gift of Horses / Reading level suggested by edHelper: Grades 7-8

Downloaded at http://edhelper.com/ReadingComprehension_29_180.html

When you hear the word “hippotherapy” what do you think of? A hippo in a hot tub? Perhaps a hippo getting a massage? It’s nothing like that at all. Hippo is the Greek word for horse. So hippotherapy is the use of horses to improve health. You may not have heard of it before, but using horses as part of physical therapy began in the U.S. and Germany late in the 1940s. It was slow to catch on, but now there are more than 600 centers for equine or hippotherapy in the United States alone. The benefits of riding horses have been known since the fifth century B.C. Those who rode horses often had better balance, muscular strength, and confidence. In spite of that, the use of horses as part of a regular therapy program has been around less than 60 years, and most of those for less than 20. What is it about riding horses that helps people? The movement of a horse’s hips as it walks is very similar to that of a person’s walk. The gentle movement you feel as you ride helps to exercise many of the same muscles used by humans, but without the effort. For those who cannot walk or who have difficulty walking, those muscles can become weak or atrophied from lack of use. Taking part in a riding therapy program works because those muscles are gently exercised. This builds strength in the

Title: Americans with Disabilities Act in 1990 / Reading level suggested by edHelper: Grades 9-12

Downloaded at http://edhelper.com/ReadingComprehension_29_182.html

America is called the "Land of Opportunity." The Statue of Liberty invites many to our country with the words: "Give me your tired, your poor, Your huddled masses yearning to breathe free, The wretched refuse of your teeming shore. Send these, the homeless, tempest-tossed to me. I lift my lamp beside the golden door." For many groups of people the battle for liberty, civil rights, and equal access has been hard fought. This includes people who are disabled. After World War I, many soldiers returned from battle permanently disabled. Prior to this war, the government granted a pension to disabled veterans. However, the help they needed to readjust to life with a disability was missing. The government stepped in to help. Now disabled veterans were given the opportunity to learn skills needed to find work and regain their daily activities. However, disabled non-veterans were still without assistance until 1935. Under the direction of Franklin D. Roosevelt (who himself was disabled due to polio), the Social Security Program was formed. This program included payments to the permanently disabled to assist them in living. After centuries of being thought of as "burdens to society," public sentiment towards the disabled began to change. However, change is difficult. The barriers in society for the disabled to overcome were tremendous. Not only were there deep-seated fears and misunderstandings in the minds of people, there were physical barriers that needed to be changed. People who used

Appendix B: Sanmin high school English textbook readings (Each of 397 words)

Title: How Does It Taste? / Reading Level: Book 1 Lesson 2

Does milk taste the same as orange juice? Of course not! Does fish taste like chicken? Not at all. But how do you know? What tells you they are different? Is it your tongue? Maybe you think so. But guess again. We do taste things with our tongues; that's true. But the smell of food has a lot to do with its taste, too. We taste foods with our noses as well as our tongues. In fact, the nose has more to do with taste then the tongue. Scientist say that your tongue can recognize only four tastes. It can tell if something is sour (like vinegar) or bitter (like soap). But that's all. To tell different foods apart, we also have to use our noses. Can you remember a time when you had a bad cold? Your food tasted very plain then. It seemed to have little taste at all. That wasn't because your tongue wasn't working. It was because your nose was stopped up. You couldn't smell the food, and that made it seem tasteless. You can prove this to yourself. Try eating something while you pinch your nose shut. It won't seem to have much taste. Here's another test. It shows how important the nose is in tasting. First you blindfold a person. Then you put a piece of potato in his mouth. You tell him to chew it. At the same time, you hold a piece of apple under his nose. Then ask what food is in his mouth. Most people will say, "An apple." The smell of the apple fools them. The test works best when two foods feel the same in the mouth. It won't work well with apple and orange slices. They don't feel alike. What about the eyes? Do they help us taste? Sometimes they may. The way a food looks can make a difference in its taste. Sometimes we taste what we expect to taste. Here's a test to show that: Get some orange food coloring. Mix some into milk. It does not change the taste. Now ask people to taste the orange milk. Ask if it tastes all right. Many people will say it tastes odd. Because it looks odd, they expect an odd taste. And so it tastes odd to them. So you see, it's not only the tongue that does the tasting!

Title: Losing Our Languages / Reading Level: Book 3 Lesson 2

The time may soon come when we say goodbye to most of the world's languages. Today humans express themselves in over 6,000 different languages, but that is quickly changing. Many experts predict that over half of these languages will disappear within the next 50 years. After 100 years, the world may use only a dozen major languages. Why? When people from different cultures live and work together much more than before, change takes place. The languages of the world's dominant cultures are replacing the languages of the smaller cultures. You're learning English right now. Could this be the beginning of the end for the Chinese language? Of course not. *Mandarin* remains the healthy, growing language at the heart of Chinese culture. Mandarin steadily continues to spread among Chinese people worldwide. Elsewhere, *Swahili* grows in Africa. Spanish continues to thrive in *South America*. *Hindi* rules

India. And of course almost everyone these days wants to learn English. However, many less common regional languages haven't been so lucky, because most young people have stopped learning them. When less common languages disappear, two factors are to blame: trade and technology. Most international trade takes place in major world languages such as English or Mandarin. Cultures that isolate themselves from international business and major world languages have difficulty prospering. Most children respect their own culture and traditions. But when it comes to getting a job, knowing a major world language if often essential. It may mean the difference between success and failure. For many, using a less common regional language simply isn't very helpful in today's world. Technology affects languages in an even more fascinating way. Modern media such as radio and television give young people in developing countries much knowledge about the world. These young people can learn about places they've never visited. Their minds open to new events and ideas. This knowledge doesn't come in words from the mouths of their parents or the elders in their community. It usually comes in the language of a dominant culture. It's not surprising then that young people are drawn away from their regional languages. Many benefits come when different cultures begin to share a common language. Instead of struggling for words, people can quickly share ideas and work together. Knowing the same language gives people from different places common ground. A shared language means easier communication and a foundation for trust.

Title: The News / Reading Level: Book 5 Lesson 1

News is an account of events that interest and concern the public. Community residents want to know about a proposed new park in town. The whole nation cares about the devastating earthquake in central Taiwan or an approaching typhoon from the Philippine Sea. To you, information about your friend's flu is news. However, not every story is newsworthy. What is news worthy in one medium may be otherwise in another. The arrival of a new teacher may be reported in the school paper but not in a national newspaper. A hotel fire may make the headlines in local newspapers but not on CNN. What makes a story newsworthy? The question may be answered with the following news elements. Unusualness: A reporter at NBC put it this way: "If an airplane departs on time, it isn't news. If it crashes, regrettably it is." In a nutshell, that comment explains news. News is the different, the unusual, and the out-of-the-ordinary. People sometimes ask, "Why is the news always bad?" Actually, most of the news media include good news, but unusual is more often found in bad news. Significance: Important events, those that affect many people, are news. Some examples are taxes, elections, wars, scientific discoveries, the economy, which are significant in people's lives. Timeliness: Old news isn't news; it's history. People want to hear about the flood while it's happening, not next month when everything has dried out. Proximity: People want to know about nearby events: burglaries in the neighborhood, the proposed regional highway, or the new income tax

law. Prominence: When well-known people, buildings, or places are involved, that is news. If you are arrested for shoplifting, it might not make even the local news. But if a movie star is arrested, that's news. Human interest: Stories about ordinary people or animals, humorous or dramatic stories, heartwarming or sad stories often appear in the news because they have human interest: an emotional and personal appeal that draws our attention. Here are two examples that can help you better understand the above concept. United States under Attack Sep 12, 2001 The United States was under attack Tuesday morning, with widespread destruction throughout the East Coast that included at least four commercial jet crashes into significant buildings. The first wave of the attack centered on the World Trade Center in *Manhattan* when a hijacked commercial airline slammed into the second

Appendix C: Lexical relations in the graded readings for American children

Target word	winter	snow	game	kid	school	cold	flu	belly	nurse	class	germ	tip	child	hand	head
Index of synset	0	1	2	0	0	0	0	0	0	1	2	2	0	0	0
Target word length	6	4	4	3	6	4	3	5	5	5	4	3	5	4	4
Number of hyponyms	N/A	2	19	53	98	1	4	2	20	3	N/A	N/A	53	9	8
Number of compounds	N/A	1	6	1	62	1	2	1	13	1	N/A	N/A	6	2	2
Hyponymous length	N/A	6	11.3	8.77	12.9	8	11.5	9.5	10.9	11.7	N/A	N/A	8.77	7	6.13
Length difference	N/A	**2**	**7.32**	5.77	**6.86**	**4**	**8.5**	**4.5**	**5.9**	**6.67**	N/A	N/A	3.77	**3**	**2.13**
Compound ratio	N/A	**50**	**31.6**	1.89	**63.3**	**100**	**50**	**50**	**65**	**33.3**	N/A	N/A	11.3	**22.2**	**25**

Wash Your Hands (Grades 1-2)

thing	diaper	food	eye	nose	mouth	door	envelope	air	sister	cabinet	package	cookie	carnival	kit	kitchen	counter
1	0	0	0	0	1	0	0	1	0	0	0	0	0	0	0	3
5	6	4	3	4	5	4	8	3	6	7	7	6	8	3	7	7
N/A	N/A	689	11	20	8	23	1	N/A	5	5	1	39	2	3	7	4
N/A	N/A	25	3	4	0	18	1	N/A	5	2	0	15	0	2	1	0
N/A	N/A	8.8	7.91	6.3	4.63	9.83	14	N/A	10.4	11.6	4	10.8	9.5	8.67	8	7.75
N/A	N/A	4.8	**4.91**	**2.3**	-0.4	**5.83**	**6**	N/A	**4.4**	**4.6**	-3	**4.77**	1.5	**5.67**	1	0.75
N/A	N/A	3.63	**27.3**	**20**	0	**78.3**	**100**	N/A	**100**	**40**	0	**38.5**	0	**66.7**	14.3	0

The MDA Carnival Package Arrives (Grades: 3-4)

glass	milk	plate	snack	chair	content	money	million	dollar	thousand	bite	gulp	dream	head	mouth	armload	grocery
1	0	3	0	0	0	0	1	0	0	1	0	2	0	0	not	1
5	4	5	5	5	7	5	7	6	8	4	4	5	4	5	found	7
21	24	7	4	48	N/A	48	N/A	25	1	11	N/A	3	8	6	N/A	2
9	24	6	0	31	N/A	1	N/A	24	0	0	N/A	0	2	0	N/A	1
8.71	11.2	10.1	8.5	10.6	N/A	8.33	N/A	13.9	9	5.18	N/A	11	6.13	4.17	N/A	12.5
3.71	**7.17**	**5.14**	3.5	**5.65**	N/A	3.33	N/A	**7.92**	1	1.18	N/A	6	**2.13**	-0.8	N/A	**5.5**
42.9	**100**	**85.7**	0	**64.6**	N/A	2.08	N/A	**96**	0	0	N/A	0	**25**	0	N/A	**50**

bag	grandmother	fall	surgery	hip	rehabilitation	hospital	therapist	car	patch	ice	leg	crutch	bit	disorder	motor	development
0	0	1	3	1	3	0	0	0	0	0	0	0	3	0	1	2
3	11	4	7	3	14	8	9	3	5	3	3	6	3	8	5	11
91	1	2	193	N/A	N/A	32	22	75	25	7	15	N/A	N/A	302	N/A	57
47	0	1	23	N/A	N/A	6	5	20	0	1	4	N/A	N/A	39	N/A	2
8.04	3	7.5	12.9	N/A	N/A	10.6	10.9	8.33	8.36	6.43	8.33	N/A	N/A	14.1	N/A	11.6
5.04	-8	**3.5**	5.9	N/A	N/A	2.59	**1.91**	**5.33**	3.36	3.43	**5.33**	N/A	N/A	6.12	N/A	0.56
51.6	0	**50**	11.9	N/A	N/A	18.8	**22.7**	**26.7**	0	14.3	**26.7**	N/A	N/A	12.9	N/A	3.51

Physical Therapists (Grades 4-6)

gym	day	week	muscle	ball	bench	swing	swing	therapy	treatment	individual	body	disability	illness	injury	movement	effect
0	0	0	0	0	0	1	3	0	0	0	0	0	0	0	1	0
3	3	4	6	4	5	5	5	7	9	10	4	10	7	6	8	6
N/A	16	8	196	44	10	1	N/A	93	174	2045	27	153	579	124	331	44
N/A	5	2	59	34	3	0	N/A	44	4	1	13	1	2	2	0	10
N/A	7.75	7.5	15.6	9.3	8.2	7	N/A	14.8	13.3	8.45	8.07	12.4	12.9	10.5	7.64	9.86
N/A	**4.75**	**3.5**	**9.58**	**5.3**	**3.2**	2	N/A	**7.83**	4.26	-1.5	**4.07**	2.37	5.86	4.48	-0.4	**3.86**
N/A	**31.3**	**25**	**30.1**	**77.3**	**30**	0	N/A	**47.3**	2.3	0.05	**48.1**	0.65	0.35	1.61	0	**22.7**

minimum	recovery	doctor	field	people	part	outbreak	world	war	school	word	hippotherapy	hippo	tub	massage	nothing	horse
0	1	0	3	0	0	0	1	0	0	0	Not	1	0	0	0	0
7	8	6	5	6	4	8	5	3	6	4	found	5	3	7	7	5
1	2	99	320	187	1076	3	3	20	98	185	N/A	N/A	4	8	5	134
0	0	11	0	19	1	0	0	5	62	34	N/A	N/A	1	3	0	42
8	5	11.5	12.4	8.84	8.96	9.67	8.33	11.8	12.9	10.5	N/A	N/A	7.25	11	9.2	8.94
1	-3	5.52	7.39	2.84	4.96	1.67	3.33	**8.75**	**6.86**	6.55	N/A	N/A	**4.25**	**4**	2.2	**3.94**
0	0	11.1	0	10.2	0.09	0	0	**25**	**63.3**	18.4	N/A	N/A	**25**	**37.5**	0	**31.3**

Gift of Horses (Grades 7-8)

use	health	part	therapy	center	equine	benefit	century	balance	strength	confidence	program	people	movement	hip	walk	muscle
0	0	0	0	2	0	1	0	0	0	0	0	0	1	0	0	0
3	6	4	7	6	6	7	7	7	8	10	7	6	8	3	4	6
65	N/A	1076	93	13	169	5	2	3	40	N/A	153	187	331	N/A	40	196
8	N/A	1	44	8	0	0	1	2	0	N/A	3	17	0	N/A	1	59
13.2	N/A	8.96	14.8	13	8.95	6.6	14	13.3	9.05	N/A	10.4	8.84	7.64	N/A	7.6	15.6
10.2	N/A	4.96	**7.83**	**7**	2.95	-0.4	**7**	**6.33**	1.05	N/A	3.42	2.84	-0.4	N/A	3.6	**9.58**
12.3	N/A	0.09	**47.3**	**61.5**	0	0	**50**	**66.7**	0	N/A	1.96	9.09	0	N/A	2.5	**30.1**

human	effort	difficulty	lack	land	opportunity	statue	liberty	country	word	mass	refuse	shore	lamp	door	group	people
0	1	0	0	0	0	0	1	0	1	5	0	0	0	2	0	0
5	6	10	4	4	11	6	7	7	4	4	6	5	4	4	5	6
27	103	2	12	8.95	40	9	8	23	N/A	12	N/A	15	68	1	723	187
0	1	0	0	31	0	1	0	3	N/A	0	N/A	2	27	1	19	19
10.6	10.1	7	9.25	12.9	8.85	8.33	7.75	10.3	N/A	8.58	N/A	7.73	10.3	8	8.76	8.84
5.56	4.15	-3	5.25	13.5	-2.2	2.33	0.75	3.35	N/A	4.58	N/A	2.73	**6.32**	**4**	3.76	2.84
0	0.97	0	0	4	0	11.1	0	13	N/A	0	N/A	13.3	**39.7**	**100**	2.63	10.2

Americans with Disabilities Act in 1990 (Grades: 9–12)

battle	right	access	world	war	soldier	government	veteran	skill	work	activity	assistance	direction	polio	program	payment	century
1	0	1	1	0	0	0	1	0	2	0	0	6	0	1	1	0
6	5	6	5	3	7	10	7	5	4	8	10	9	5	7	7	7
5	109	N/A	3	20	71	36	1	24	22	993	89	15	N/A	18	21	2
0	34	N/A	0	5	2	9	0	0	3	5	0	0	N/A	13	4	1
5.6	13.7	N/A	8.33	11.8	8.79	12.6	11	10.4	9.05	9.62	9.55	12.8	N/A	15.4	10.4	14
-0.4	**8.74**	N/A	3.33	**8.75**	1.79	**2.58**	4	5.38	5.05	1.62	-0.4	3.8	N/A	**8.44**	3.43	**7**
0	**31.2**	N/A	0	**25**	2.82	**25**	0	0	13.6	0.5	0	0	N/A	**72.2**	19	**50**

burden	society	sentiment	change	barrier	fear	misunderstanding	mind
0	0	1	3	1	0	1	2
6	7	9	6	7	4	16	4
4	35	18	2	9	49	N/A	4
0	3	0	0	2	0	N/A	0
7.5	15.8	10.8	14	11.1	8.35	N/A	11.8
1.5	8.8	1.83	8	**4.11**	4.35	N/A	7.75
0	8.57	0	0	**22.2**	0	N/A	0

Appendix D: Lexical relations in the selected readings of Taiwanese high school English textbooks

Target word	milk	orange	juice	fish	chicken	tongue	thing	smell	food	taste	nose	scientist	vinegar	soap	time
Index of synset	0	0	0	1	0	0	7	0	0	0	0	0	0	0	6
Target word length	4	6	5	4	7	6	5	5	4	5	4	9	7	4	4
Number of hyponyms	24	8	8	50	10	N/A	11	18	689	29	20	164	3	14	64
Number of compounds	24	8	6	2	0	N/A	0	0	25	0	4	13	3	12	10
Hyponymous length	11.2	11.8	9.13	9.1	7.7	N/A	8.45	5.94	8.8	7.21	6.3	12.4	11.7	10.1	8.11
Length difference	**7.17**	**5.75**	**4.13**	5.1	0.7	N/A	3.45	0.94	4.8	2.21	**2.3**	3.37	**4.67**	**6.07**	4.11
Compound ratio	**100**	**100**	**75**	4	0	N/A	0	0	3.63	0	**20**	7.93	**100**	**85.7**	15.6

How Does It Taste? (Book 1 Lesson 2)

cold	test	person	piece	potato	mouth	apple	people	slice	eye	way	difference	coloring	time	goodbye	language	human
0	1	0	7	0	0	0	0	1	0	0	0	0	2	0	0	0
4	4	6	5	6	5	5	6	5	3	3	10	8	4	7	8	5
1	26	2045	8	10	6	29	187	8	11	38	43	N/A	44	N/A	342	27
1	8	75	0	4	0	5	19	0	3	0	0	N/A	0	N/A	91	5
8	16.5	8.45	7.25	11.6	4.17	10.5	8.84	7.25	7.91	9.34	10.4	N/A	7.93	N/A	11.2	10.6
4	**12.5**	2.45	2.25	**5.6**	-0.8	5.48	2.84	2.25	**4.91**	6.34	0.37	N/A	3.93	N/A	**3.22**	5.56
100	**30.8**	3.67	0	**40**	0	17.2	10.2	0	**27.3**	0	0	N/A	0	N/A	**26.6**	18.5

Losing Our

expert	year	world	dozen	people	culture	change	beginning	end	heart	day	factor	trade	technology	business	difficulty	child
0	0	1	0	0	0	2	0	5	4	1	0	0	0	1	0	0
6	4	5	5	6	7	6	9	3	5	3	6	5	10	8	10	5
260	22	3	1	187	24	764	84	N/A	6	14	10	5	18	160	2	53
5	13	0	0	19	11	10	0	N/A	0	4	3	3	6	6	0	6
10.3	9.09	8.33	7	8.84	17.1	9.2	10.1	N/A	8.8	11	12	13	16	11	7	8.8
4.32	**5.09**	3.33	2	2.84	**10.1**	3.2	1.06	N/A	3.8	**7.8**	**6.4**	**7.6**	**6.1**	2.8	-3	3.8
1.92	**59.1**	0	0	10.2	**45.8**	1.31	0	N/A	0	**29**	**30**	**60**	**33**	3.8	0	11

Languages (Book 3 Lesson 2)

tradition	job	difference	success	failure	way	medium	radio	television	country	knowledge	place	mind	event	idea	word	parent
0	0	0	2	2	0	0	0	0	0	0	0	0	0	0	1	0
9	3	10	7	7	3	6	5	10	7	9	5	4	5	4	4	6
N/A	253	43	2	4	38	102	N/A	4	23	749	93	7	1085	432	N/A	40
N/A	1	0	0	2	0	4	N/A	2	3	9	8	2	9	3	N/A	4
N/A	10	10	7	10	9.3	10	N/A	12	10	9.7	8.62	10.4	9.2	10	N/A	7.1
N/A	7.3	0.4	0	**3**	6.3	4.4	N/A	**2.3**	3.3	0.7	3.62	**6.43**	4.2	6	N/A	1.1
N/A	0.4	0	0	**50**	0	3.9	N/A	50	13	1.2	8.6	**28.6**	0.8	0.7	N/A	10

mouth	elder	community	benefit	ground	communication	foundation	trust	news	account	event	community	resident	park	town	nation	earthquake
1	0	0	1	3	0	0	2	1	0	0	0	0	0	0	1	0
5	5	9	7	6	13	10	5	4	7	5	9	8	4	4	6	10
8	3	23	5	1	171	5	2	30	22	1085	23	18	2	8	17	9
0	0	1	0	1	6	0	0	11	0	9	1	0	2	6	0	1
4.6	5.3	8.3	6.6	12	10	8	11	9	11	9.2	8.3	8.1	11	7.9	8.5	11
-0	0.3	-1	-0	**6**	-3	-2	6	**5**	4.2	4.2	-1	0.1	**7**	**3.9**	2.5	0.7
0	0	4.3	0	**100**	3.5	0	0	**37**	0	0.8	4.3	0	**100**	**75**	0	11

The News (Book 5 Lesson 1)

typhoon	sea	information	friend	flu	story	medium	arrival	teacher	school	newspaper	hotel	fire	headline	question	element	unusualness
0	0	0	0	0	4	0	0	0	0	0	0	0	0	0	0	0
7	3	11	6	3	5	6	7	7	6	9	5	4	8	8	7	11
N/A	1	250	37	4	15	102	3	40	98	7	30	17	9	N/A	17	19
N/A	0	5	3	2	0	4	0	12	62	1	3	12	0	N/A	1	0
N/A	8	9.8	7.4	12	9.6	10	7.3	11	13	7.6	8.2	8.8	9	N/A	12	9.7
N/A	5	-1	1.4	**8.5**	4.6	4.4	0.3	**3.9**	**6.9**	-1	3.2	**4.8**	1	N/A	4.5	-1
N/A	0	2	8.1	**50**	0	3.9	0	**30**	**63**	14	10	**71**	0	N/A	5.9	0

reporter	airplane	time	nutshell	comment	significance	example	tax	election	war	discovery	economy	life	timeliness	history	flood	month
0	0	6	0	0	0	0	0	0	0	1	0	1	0	0	0	0
8	8	4	8	7	12	7	3	8	3	9	7	4	10	7	5	5
8	55	64	N/A	35	9	13	64	10	20	N/A	29	N/A	N/A	N/A	8	111
2	2	10	N/A	1	0	0	32	5	5	N/A	8	N/A	N/A	N/A	4	5
12	12	8.1	N/A	7.6	10	8.8	9.8	11	12	N/A	13	N/A	N/A	N/A	10.5	7.02
3.6	3.6	4.1	N/A	0.6	-2	1.8	**6.8**	**3.3**	**8.8**	N/A	**5.9**	N/A	N/A	N/A	**5.5**	2.02
25	3.6	16	N/A	2.9	0	0	**50**	**50**	**25**	N/A	**28**	N/A	N/A	N/A	**50**	4.5

proximity	burglary	neighborhood	highway	income	law	prominence	building	place	shoplifting	movie	star	interest	animal	appeal	attention	concept
0	0	0	0	0	1	0	0	0	0	0	1	0	0	1	2	0
9	8	12	7	6	3	10	8	5	11	5	4	8	6	6	9	7
N/A	4	9	23	94	19	11	351	93	N/A	37	1	19	393	3	2	380
N/A	0	0	4	9	11	0	10	8	N/A	2	1	0	11	0	0	4
N/A	13	7.22	9.43	10	12.9	9.27	9.5	8.6	N/A	9.7	9	9.2	8.5	9.7	6	12
N/A	5	-4.8	2.43	4.04	**9.89**	-0.7	1.5	3.6	N/A	4.7	**5**	1.2	2.5	3.7	-3	4.9
N/A	0	0	17.4	9.57	**57.9**	0	2.8	8.6	N/A	5.4	**100**	0	2.8	0	0	1.1

state	attack	morning	destruction	coast	jet	crash	wave	world	trade	center	airline
0	0	0	1	0	0	1	1	1	0	2	1
5	6	7	11	5	3	5	4	5	5	6	7
9	105	N/A	17	8	8	1	N/A	3	5	13	N/A
4	8	N/A	0	0	7	0	N/A	0	3	8	N/A
13	11.4	N/A	9.12	7.88	7.5	5	N/A	8.33	12.6	13	N/A
8.1	5.37	N/A	-1.9	2.88	**4.5**	0	N/A	3.33	**7.6**	**7**	N/A
44	7.62	N/A	0	0	**87.5**	0	N/A	0	**60**	**61.5**	N/A

Computational Linguistics and Chinese Language Processing
Vol. 14, No. 1, March 2009, pp. 85-104

Summarization Assistant for News Brief Services on Cellular Phones

Yuen-Hsien Tseng*

Abstract

A Chinese news summarization method is proposed in order to help humans deal with the message services of news briefs broadcast over cell phones. The problem to be solved here is unique because a strict length limit (69 or 45 characters) is imposed on the summaries for the message service. This requires some sort of automatic sentence fusion, rather than sentence selection alone. In the proposed method, important sentences were first identified based on the news content. They were matched against the news headline to determine a suitable position for concatenation with the headline to become candidates. These candidates were then ranked by their length and fitness for manual selection. In our evaluation, among 40 short news updates in the inside testing set, over 75% (80%) of the best candidates yield acceptable summaries without manual editing for the length limit of 69 (45) characters. These numbers, however, reduce to 70.7% (53.3%) for the outside testing set of 75 news stories of ordinary length. It seems that the shorter the length limit, the more difficult the problem of getting the summary from long stories. Nevertheless, the proposed method has the potential not only to reduce the cost of manual operation, but also to integrate and synchronize with other media in such services in the future.

Keywords: Cell Phone Service, News Brief Message, Automated Summarization, Chinese News

1. Introduction

The popularity of cell phones in the Taiwan area has reached the highest rate in the world during the last few years. Over 23 million cell phone numbers were used as of June 2002, which is slightly more than the population of Taiwan (Wang, 2002). To better utilize this ubiquitous communication device, a number of content providers have provided Chinese news

* National Taiwan Normal University, No.162, Sec. 1, Heping East Road., Taipei City, Taiwan (R.O.C.), 106, Tel: 886-2-77345535
E-mail: samtseng@ntnu.edu.tw

[Received February 13, 2009; Revised July 16, 2009; Accepted July 17, 2009]

brief services over the cell phone, such as United Daily News (United Daily News, n.d.), Central News Agency (Central News Agency, n.d.), and PC Home in Taiwan. The Asahi Shimbun in Japan (the second largest news agency in the world) has provided such news message services at an inexpensive rate since 1999, in the hope that the increase in the number of the readers of their content could lead to an increase in the subscriptions to their newspaper (China Times, n.d.). As multimedia technologies continue to improve, future news service over the cell phone may not only include text, but also speech, images, or video, integrated and synchronized. To reach this vision, however, the operation cost should be low enough to sustain such services. Therefore, automated methods of cost containment would be of great help.

The news brief shown on a cell phone is different from one on a desktop computer. Due to the limited screen size, a length limit is defined for each news message. This is usually 45 Chinese characters in PHS systems or 69 characters in other systems, including punctuation marks (United Daily News, n.d.). Summaries of this kind are longer than a news headline but shorter than a long Chinese sentence. For the benefit of the subscribers, the summaries should contain as much content as possible to reduce the frequency of retrieving the whole news story. Also the readability and coherence of the summaries are important factors that should be taken into account.

From the research perspective, the task defined above is a challenge for automatic document summarization. Previous studies have shown that the shorter the summary required, the lower the performance of machine-generated summary (Lin & Hovy, 2003), hence, the more difficult the problem is. The task of news brief summarization for cell phones falls into this difficult category. On the other hand, human summarization of news stories for cell phones is not really a difficult problem. As mentioned above, the main issue is whether one can achieve this task in a low-cost and efficient way. Strictly maintaining the length limit requires a human summarizer to pay attention to the number of characters already there while making the summarization. If a machine could suggest a number of summary candidates, each with its length shown, for human selection, not only would the human summarizer be relieved of such tedious work and improve his/her efficiency, but also the task would become less difficult for machine summarization.

This article proposes a Chinese news summarization technique to assist human summarizers in the above way, with the aim of meeting the considerations described above. Basically, our approach is a sentence fusion technique that merges the news headline with the body sentence that supplements the information carried by the headline. After a brief review of previous work in the next section, the detailed approach and its motivations are described. The performance is then evaluated and the results are shown. This is followed by a discussion of the strengths and weaknesses of the proposed method. Finally, we conclude this paper with

some other possible applications and future work for further exploring Chinese news summarization techniques.

2. Related Work

Automatic news summarization techniques have been widely explored in recent years, such as the summarization tasks in DUC (DUC, n.d.) or in NTCIR (Fukusima, Okumura, & Nanba, 2002). Several practical systems (e.g. (Hovy & Lin, 1999; Evans, Klavans, & McKeown, 2004; Radev, Otterbacher, Winkel, & Blair-Goldensohn, 2005)) have been developed in the past decade. The summarization techniques used in most studies can be divided into two approaches: abstraction and extraction (Mani, 2001; Radev, Hovy & McKeown, 2002). In abstraction, advanced natural language processing (NLP) techniques are applied to analyze sentential information and then to generate concise sentences with proper semantics. Sophisticated NLP techniques, such as anaphora resolution, may be used and certain human maintained knowledge bases or corpora may be needed. In extraction, statistical techniques are applied to rank and select the text snippets for a summary. Due to its relatively low cost and high robustness across application domains and document genres, most summarization tasks adopt the extraction approach (Carbonell & Goldstein, 1998; Lin & Hovy, 2002, Tseng, *et al*, 2007). Nevertheless, abstraction-based methods move the summarization field from the use of purely extractive methods to the generation of abstracts that contain sentences not found in any of the input documents and also synthesize information across sources (Barzilay & McKeown, 2005). Thus, the need for an abstraction-based approach is sometimes inevitable.

Despite the vast literature already published, most of the studies are for English. Although some have focused on Chinese news (e.g. (Chen, Kuo, Huang, Lin & Wung, 2003)), none have been done for the problem discussed here. The problem to be solved in this paper is unique due to the facts that there is a strict length limit imposed and that the range of the length limit makes most simple sentence selection approaches invalid. Thus, an abstraction-based method or a similar one that requires sentence fusion or alteration is required.

For example, in (Takefumi, Hidetaka, & Hiroshi, 2003) the authors reported a deletion-based approach to summarize a Web news article for PC to another short article for cell phones for Japanese. There, the length limit of the short article ranges from 50 to 100 Japanese characters. The approach first computes the values of TFxIDF for each clause in advance. A few significant sentences from the original article are then extracted based on the TFxIDF values. After that, verbose descriptions corresponding to the leaves of the dependency trees, having the lowest TFxIDF, are removed from the sentences until the length of the result of summarization is within the limit.

An important issue in automatic summarization is the evaluation of machine-derived summaries. This is not an easy task. Two main approaches are commonly applied: intrinsic and extrinsic evaluation (Mani, 2001). In intrinsic evaluation, manually prepared answers or evaluation criteria are compared with those that are machine generated. In extrinsic evaluation, automated summaries are evaluated based on their performance or influence on other tasks, such as document categorization. We adopt the intrinsic approach here since it is obviously suitable for our task.

3. The Proposed Summarizer

To develop an automated Chinese news summarizer subject to the limitations of a cell phone, an understanding of the style of the news stories and how humans summarize them would be helpful. Table 1 lists three news examples and their English translations. As can be seen, these examples are short, with their bodies having only 1, 2, and 3 sentences, respectively. This is not uncommon for the stories to be transmitted to users' cell phones, although longer stories may be selected as well. Given such short stories, a human summarizer has very few clues as to rewrite the story thoroughly to fit the length limit. The best he or she can do may be to cut and paste the snippets from the news text with minimal editing to avoid garbling the original meaning.

The snippets to be cut and pasted can be enumerated then suggested by a computer for manual selection. Nevertheless, the possibilities of such enumeration would be huge if all substrings of the news text are blindly considered. As can be seen from the examples in Table 1, a Chinese sentence is often composed of several comma-separated clauses, which convey the meaning of the sentence in successive sequence. Chinese clauses are independent from each other in some circumstances and, thus, constitute a useful unit to be combined with others to make a new sentence. Although most of the combined sentences would be invalid, several of them would still be meaningful and sometimes more complete in content, especially for those from the beginning and ending clauses.

Table 1. Three news examples for summarization[1]. The number in parenthesis is the number of characters in the preceding sentence.

1	太空探測器在遙遠的恒星周圍發現水的痕跡 (19) 美國航空航太總署的科學家星期三稱，新近在一顆遙遠的恒星周圍發現了水存在的痕跡，這可以成爲第一個支援除我們自己存在地外生命的證據。(64) #2001/07/13#
	Space Probe Sees Signs of Water Around Distant Star Newly detected signs of water around a distant star are the first evidence that planetary

[1] The first two stories were accessed on 2005/01/04 from http://www.1999.com.tw/english/, while the third story was accessed on 2005/01/05 at http://news2.ngo.org.tw/php/ens.php?id=03102302

	systems outside our own might be able to support life, NASA scientists said on Wednesday.
2	**專家：世界人口接近頂點 90 億** (13) 科學家星期三預測說，在 2070 年左右，世界人口可能會達到頂峰約 90 億，然後開始下降。(38) 澳大利亞人口統計學家在考慮很多因素後計算出到本世紀結束時，地球上的人口會下降至 84 億人。(43) #2001/08/03#
	Experts: World Population Set to Peak at 9 Billion The world's population will probably peak at about 9 billion around 2070 before it starts to decline, scientists predicted Wednesday. Demographers at a think tank in Austria calculate that by the turn of the century the number of people on the planet will have dropped down to 8.4 billion people.
3	**海洋生物普查行動發現數百種新生物** (16) 有史以來第一次的「海洋生物普查計畫」進行以來，來自 53 個國家的科學家們，平均每星期便發現 3 種新的海洋魚種。(52) 這項爲期 10 年、耗資 10 億美元的計畫，動員了來自世界各地的科學家共同合作，目的是爲了將存在海洋中所有種類的生物發掘出來並予以分類。(63) 計畫至今已實施了 3 年，科學家在海裡發現了 15300 多種生物，他們估計還有 5 千多種生物尚未被科學界發掘。(49) #2003/10/23#
	Marine Life Census Finds Hundreds of New Species New marine fish species are being logged at an average rate of three per week by scientists from 53 countries engaged in the first Census of Marine Life. The 10 year, $1 billion global scientific collaboration aims to identify and catalog all life in the oceans. After their first three years of work, census scientists report over 15,300 species of fish in the sea and estimate 5,000 more are still unknown to science.

Table 2. The summary candidates for the third story in Table 1. They were created by combining the last clause of each body sentence with the headline.

1	海洋生物普查行動發現數百種新生物，平均每星期便發現 3 種新的海洋魚種。(34) Marine life census finds hundreds of new species, at an average rate of three per week.
2	海洋生物普查行動發現數百種新生物，目的是爲了將存在海洋中所有種類的生物發掘出來並予以分類。(45) Marine life census finds hundreds of new species, aims to identify and catalog all life in the oceans.
3	海洋生物普查行動發現數百種新生物，他們估計還有 5 千多種生物尚未被科學界發掘。(38) Marine life census finds hundreds of new species; they estimate 5,000 more are still unknown to science.

Take the third story from Table 1 as an example. The headline has only 16 characters, falling short of the required length of 45 or 69. The other 3 sentences have 52, 63, and 49 characters, respectively. None of them alone is an ideal summary of the required length. Nevertheless, *by concatenating the headline and the last clause of each body sentence*, as shown in Table 2, each becomes a better choice for summaries of length 45.

It is noted that the simple Select-First-N strategy which usually has been the baseline method for most news summarization tasks would not work here. As can be seen from Table 1, if the first *n* characters were used as the summary, the underlined text in the first sentence of the story would be chosen as the summary for the length limit 45. These summaries, however, are incomplete in their meaning. If the first clauses (ending with a comma or period and no longer than *n*) were used as the summary, they may be still incomplete in meaning or too verbose to deliver the message even when their meanings are complete. For example, for the length limit 45, the first two clauses of the first story: "美國航空航太總署的科學家星期三稱，新近在一顆遙遠的恒星周圍發現了水存在的痕跡" (38 characters) would be extracted based on the Select-First-N clause strategy. This, however, is inferior to the perfect summary "太空探測器在遙遠的恒星周圍發現水的痕跡，這可以成爲第一個支援除我們自己存在地外生命的證據。" (45 characters) which is extracted based on the heuristic rule shown in Table 2. In this example, the headline perfectly replaces the first two clauses, leaving more space for including the supplemental information in the final clause: "這可以成爲第一個支援除我們自己存在地外生命的證據。".

The above observation gives us clues to effectively enumerate the summary candidates. Nevertheless, there are other problems that need to be considered in order to further reduce the burden of human selection: (1) The number of suggested candidates should be fairly equal for each story. Long stories should not yield considerably more candidates than short ones. (2) The candidates should be ranked in some sense when they are suggested for selection.

To tackle these problems, we propose the following processing steps:

Step 1: Sort all the sentences of a news story by their weights and select the best 5 sentences for use in the next step.

Step 2: Generate summary candidates by matching and combining each selected sentence with the news headline. Calculate the match scores and summary lengths.

Step 3: Sort the candidates by their lengths and scores.

In Step 1, the weight of a sentence in a story of any length is determined by the accumulated weights of the keywords occurring in that sentence, as shown below:

$$weight(S) = \sum_{w \in Keywords \in S} (0.5 + 0.5 \times tf_w / \max\_tf)$$

where tf_w is the term frequency of keyword *w* and *max_tf* is the term frequency of the keyword which occurs most in the news story. Here, the keywords of a story are those headline words that remain from non-content-bearing word deletion and those maximally repeated patterns in the story that are extracted by Tseng's algorithm (Tseng, 2002). Tseng has shown that Chinese news stories can contain many new keywords, almost 1/3 of repeated words are unknown to a lexicon of 123,226 terms. His algorithm ensures that unknown words can be extracted as well, as long as they occur at least twice in a document.

In Step 2, since headlines are guides to a news story, they should be included in the beginning of the candidates. The ending clauses to be concatenated should supplement the content of the headline. This means that the beginning clauses of a body sentence should be as similar to the headline as possible. To spot the position for concatenation and to know the similarity, a dynamic programming (DP) technique is used.

Given two strings A[1..n] and B[1..m], where n<=m, the edit distance between A[1..i] and B[1..j] based on DP (Levenshtein, 1966) is:

$$d[i, j] = min(d[i-1, j] , d[i-1, j-1], d[i, j-1])+ c(A[i], B[j])$$

where *min* is a function that returns the minimum of its 3 arguments, and c(A[i], B[j]) = 0 if A[i]=B[i], and 1 otherwise. The initial values for the distance are: d[0, 0]=0, d[0, j]=0 for j=1..m and d[i, 0] = d[i-1,0]+1 for i=1..n.

A similarity function is defined in (Lopresti & Zhou, 1996) to convert the edit distance into the similarity: *exp*(d[n,j] / (d[n,j] - n)), where *exp* is the exponent function. This similarity ranges from 0 to 1. We found, however, that its range does not distribute well for later comparison. Thus, it was changed into:

$$sim(j) = \exp(\frac{d[n, j]}{d[n, j] - m - n})$$

where j denotes the j-th character (including the punctuation) of the body sentence. The new measure ranges from exp(-n/m) to 1.

The starting position (the position for concatenation) of the ending clauses is first determined by the comma which most closes the character with highest similarity. Since we favor length more than similarity (here, length is a direct measure that must be met, while similarity is just an approximation of the content similarity between the headline and the body sentence), the starting position is changed to its preceding or succeeding comma whenever such changes fit the length limit better.

Figure 1 shows an example where the second row beneath the body sentence indicates an assumed similarity score for each character position. Although the last comma (defining the starting position of the proper ending clause) has a similarity score 0.5, higher than the one

with 0.25, the desired ending clauses would start from the one with 0.25 since it fits the length limit better. This changes the summary candidate from "最好的一句，其實在最後一小句。", with 15 characters in length and 0.5 in similarity, into "最好的一句，看了以後，其實在最後一小句。", with 20 characters in length and 0.25 in similarity.

Title:

最	好	的	一	句

Body Sentence:

…	，	看	了	以	後	，	其	實	在	最	後	一	小	句	。
…	0.25	0.3	0.35	0.4	0.45	0.5	0.55	0.6	0.65	…	…	…	…	…	

Figure 1. An example to show the need for changing the starting position for better ending clauses.

Sometimes, the last clause of each body sentence may be too long, causing all the concatenated candidates to exceed the required length. In this case, only the news headline will be generated as the output without any concatenation.

News stories are often written in a so-called pyramid style where the later the paragraph occurs, the more details it carries. Thus, better summaries often come from the first few sentences. Therefore, in our current implementation, we decrease the similarity of the summary candidates composed from the sentences other than the first two by a factor of 0.85, if the number of sentences in the story exceeds *3*.

In Step 3, the length of the summary candidate is divided by the required length limit (45 or 69) to yield the length ratio ranging from 0 to 1. Now, we come to a problem of determining the rank of these candidates based on their length ratios and similarities. Ideally, this problem can be solved by machine learning methods, but they require manually prepared data to train a classifier to determine the best or to rank the candidates. The effectiveness of such machine classifiers depends heavily on the amount of training data. Since sufficient training data are difficult to prepare, a set of hand-crafted rules are devised instead:

(1) From the candidate list, find the candidate with highest similarity, called X, and the candidate with largest length ratio, called Y.

  (i) If $X = Y$, then output X.

  (ii) If $sim(X) > 1.25 * sim(Y)$ and $ratio(X) > 0.75 * ratio(Y)$, then output X, otherwise output Y.

  (iii) Remove the candidate just output from the list.

(2) Repeat (1) until there is no candidate.

4. Performance Evaluation

Based on the above steps, the summary candidates of length limit 45 for the third story in Table 1 are exactly the three in Table 2. The (ratio, similarity) values for the candidates are (0.7556, 0.7351), (1.0, 0.7757), and (0.8444, 0.6718), respectively. Step 3 sorts Candidates 1, 2, and 3 into 2, 3, and 1 in decreasing order of rank. As to the quality of the candidates, Candidate 2 with rank 1 is correct and coherent in meaning and is perfect in length. Candidate 3 is fair in Chinese expressions. It would become better if the word: "他們" ("they") in the beginning of the second clause were deleted. Candidate 1 is also correct and coherent. It carries more interesting content than Candidate 2, but it is shorter in length.

To further evaluate the above method, two sets of news stories were used. One set contained 40 Chinese news stories, which were real-time news (short stories updated every 30 minutes) from China Times[2] between August and September in 2003. Some of the stories, together with the above examples, were used to tune the parameters mentioned in the previous section. Therefore, this group of stories can be considered an inside testing set. The other set contains 75 normal stories, also from China Times between April 4 and 12 in 2009. The parameters and programs set for the inside testing set were used for these recent stories. Thus, they can be considered our outside testing set. Table 3 shows some statistics about these stories in the two sets.

Table 3. Statistics of the news stories in the two testing sets. (a) Various averages. (b) Number of documents in each category.

Average Statistics	Inside testing set	Outside testing set
Average number of sentences per story	2.95	8.51
Average number of clauses per sentence	4.08	3.97
Average number of characters per sentence	64.54	52.54
Average number of characters per clause	15.83	13.22
Average number of characters per headline	16.63	14.87

Category	No. of Doc. in inside testing set	No. of Doc. in outside testing set
政治/politics	2	10
社會/criminals	0	10
財經/economics	14	10
國際/international affairs	6	10
科技/technology	4	10
娛樂/entertainment	10	10
運動/sports	4	15

[2] "China Times" http://www.chinatimes.com.tw/

For each story in both sets, summary candidates were generated and ranked by the proposed method. A human summarizer chose a candidate that he/she thought to be the best among the candidates. The chosen one was then labeled in terms of its quality with one of the three tags: G (good), F (fair), or B (bad) if it was correct and coherent, correct with some readability, or unacceptable, respectively. The inside testing set was evaluated by one human summarizer, while the outside testing set was evaluated by 15 people each for 5 stories. All of the evaluators majored in library science, thus, have some sense of knowledge for manual summarization.

Table 4 shows the results for the inside testing set, where for each news story the title and the best candidates for length limit 45 and 69 are shown, respectively. The actual lengths of the best candidates are shown in the second column. In the fourth column, with a *, the number of body sentences in that story is shown in the title row, while the rank of the chosen candidate is listed besides the candidate. The last column indicates the manual judgment of the machine-generated summary.

Table 4. The forty news headlines, their machine-generated summaries, and manual judgment of quality for the inside testing set.

ID	Content		*	**
1	Title	台鐵計軸器採購下周進行第 11 度招標	2	
	45	台鐵計軸器採購下周進行第 11 度招標，擁有這項產品製造技術的歐洲廠商，已摩拳擦掌準備進場搶標。	1	G
	66	台鐵計軸器採購下周進行第 11 度招標，不限定廠商使用材質，下周公告招標後，等標期約 28 天、審查作業 10 天，最快 10 月中旬可以最低價格進行決標。	1	G
2	Title	台十一線濱海公路山崩，交通中斷	5	
	45	台十一線濱海公路山崩，交通中斷，造成豐濱鄉對外交通完全中斷，民眾必須往台東縣才能找到出路。	1	G
	69	台十一線濱海公路山崩，交通中斷，形成九十度丁坡度，連日來花蓮間歇性豪雨不斷，該地段今天早上九點多終於發生小規模山崩，交通中斷阻斷來往車輛。	1	G
3	Title	台鐵與工會最後協商無交集，中秋是否停駛各說各話	4	
	45	台鐵與工會最後協商無交集，中秋是否停駛各說各話，會員現在也不敢說不上班，只是應付一下主管。	1	G
	61	台鐵與工會最後協商無交集，中秋是否停駛各說各話，工會說，這是台鐵當局的一貫技倆，會員現在也不敢說不上班，只是應付一下主管。	2	G
4	Title	兩岸航空業邁進實質合作時代	1	
	43	兩岸航空業邁進實質合作時代，這項合作也正式宣布兩岸航空貨運開始走入實質合作的經營時代。	1	G
	69	兩岸航空業邁進實質合作時代，將再度齊聚廈門，出席這項兩岸航空業界首度合資的盛會，這項合作也正式宣布兩岸航空貨運開始走入實質合作的經營時代。	1	B
5	Title	高市招商，力邀重量級企業與會	2	
	30	高市招商，力邀重量級企業與會，以及多功能經貿園區的未來遠景。	1	B
	57	高市招商，力邀重量級企業與會，而行程中必定會談到世界大港高雄港和小港機場的	1	G

		海空優勢，以及多功能經貿園區的未來遠景。		
6	Title	雲縣規劃產業聚落，建立招商網路	2	
	32	雲縣規劃產業聚落，建立招商網路，發展各專區內互補特性，相互支援。	1	G
	67	雲縣規劃產業聚落，建立招商網路，並規劃以麥寮自由港區、中科雲林基地及雲林科技工業區發展爲三個相互支援發展的產業聚落，爭取更多企業投資。	1	G
7	Title	中油調高桶裝瓦斯價格	4	
	34	中油調高桶裝瓦斯價格，以二十公斤裝桶裝瓦斯來看，每桶批售價調高八元。	1	G
	64	中油調高桶裝瓦斯價格，爲反應進口成本上漲壓力，中油決定自四日零時起調漲各類液化石油氣產品牌價，調整幅度爲二・六五％至三・九四％。	1	G
8	Title	經濟部：攤販不會就地合法	1	
	29	經濟部：攤販不會就地合法，因此不會有「就地合法」這個問題。	1	B
	54	經濟部：攤販不會就地合法，未來攤販仍須先通過地方政府審核後才能獲得營業許可，因此不會有「就地合法」這個問題。	1	G
9	Title	獅、象四連戰第二役，統一獅將派出威森掛帥	3	
	43	獅、象四連戰第二役，統一獅將派出威森掛帥，親自派遣場務人員前來台北，爲威森整理投手丘。	1	G
	68	獅、象四連戰第二役，統一獅將派出威森掛帥，爭取今天晚間的勝利，統一特別從台南帶著「土坯」前來新莊，賽前將由工作人員親自爲威森整理投手丘。	1	G
10	Title	中華職棒大聯盟，教練護盤，「劉」住勝果	5	
	42	中華職棒大聯盟，教練護盤，「劉」住勝果，戰績繼續保持第一，領先獅隊的勝差拉開爲 1.5 場。	1	F
	69	中華職棒大聯盟，教練護盤，「劉」住勝果，順利終結獅隊最後反撲，拿下 1 次救援成功，距離上次（2000 年 9 月 23 日對牛隊）贏得救援成功，已將近 3 年了。	1	F
11	Title	美國網球公開賽：阿格西驚險闖進 8 強	4	
	33	美國網球公開賽：阿格西驚險闖進 8 強。阿格西遇險，險遭丹特襲擊成功。	2	G
	65	美國網球公開賽：阿格西驚險闖進 8 強；西哥畢竟老江湖，第 2 盤穩中求勝，第 3 盤守住丹特強力攻勢，終於讓小老弟因強攻不破，右腳傷重退賽。	1	G
12	Title	娜姐送吻，小甜甜人氣下滑，克莉絲汀變旺	5	
	40	娜姐送吻，小甜甜人氣下滑，克莉絲汀變旺，克莉絲汀是「一吻成名」，一夕間躍升榜首。	3	G
	50	舌吻事件這兩天在網路上引爆熱烈討論，雖然布蘭妮、克莉絲汀都被娜姐送上香吻，但人氣指數卻呈現兩個極端。	4	G
13	Title	佘詩曼睡一睡，溫碧霞脫一脫，數百萬入袋	6	
	41	佘詩曼睡一睡，溫碧霞脫一脫，數百萬入袋；而溫碧霞則是小脫一下，就賺到四百多萬台幣。	1	F
	67	佘詩曼睡一睡，溫碧霞脫一脫，數百萬入袋，最近港星佘詩曼自稱在床上睡一睡，就有六百萬台幣入袋；而溫碧霞則是小脫一下，就賺到四百多萬台幣。	1	G
14	Title	王識賢求婚很靦腆，張鳳書當老師	3	
	42	王識賢求婚很靦腆，張鳳書當老師，反倒是張鳳書教他，求婚就該在大庭廣眾下告白才有誠意。	2	F
	68	王識賢求婚很靦腆，張鳳書當老師，導演要求他下跪求婚，王識賢靦腆的說人太多，不好意思，反倒是張鳳書教他，求婚就該在大庭廣眾下告白才有誠意。	1	G

15	Title	百慕達銀行在日本開設辦事處	2	
	13	百慕達銀行在日本開設辦事處	1	B
	64	百慕達銀行在日本開設辦事處。Bermuda Global Fund Services Limited 東京辦事處將坐落于東京，並將作爲百慕達銀行旗下全球範圍的 GFS 部門與其日本客戶之間的聯繫機構。	1	G
16	Title	東芝公司同意在系統單晶片中使用 ARM 晶片	3	
	45	東芝公司同意在系統單晶片中使用 ARM 晶片，雙方已經通過新的授權協議拓展了彼此間的戰略合作關系。	1	G
	69	東芝公司同意在系統單晶片中使用 ARM 晶片，東芝公司已經同意把 ARM1026EJ-S(TM)晶片用于促成創新的系統單晶片(SOC)應用產品，從而豐富其新一代數碼產品組合。	1	G
17	Title	Inno Micro 在日本經銷並出售 nStor 產品	2	
	31	Inno Micro 在日本經銷並出售 nStor 產品，在日本出售和經銷 nStor 全系列存儲產品。	1	B
	54	Inno Micro 在日本經銷並出售 nStor 產品，日本一家私營整合商和經銷商 Inno Micro 已簽署一份協議，在日本出售和經銷 nStor 全系列存儲產品。	1	F
18	Title	登記列管繳稅營業，攤販將全面合法	2	
	34	登記列管繳稅營業，攤販將全面合法，預估有數十萬攤販可望就地「合法」。	1	G
	64	登記列管繳稅營業，攤販將全面合法，將把全台灣既存和未來可能新增的攤販，全部改以登記制統一管理，預估有數十萬攤販可望就地「合法」。	1	F
19	Title	行動攤販車可在風景區營業	4	
	41	行動攤販車可在風景區營業，甚至還成立加盟總部，鼓勵民眾只要投資數十萬元就可以創業。	1	G
	56	行動攤販車可在風景區營業，包括行動咖啡館、行動彩印店等，甚至還成立加盟總部，鼓勵民眾只要投資數十萬元就可以創業。	1	F
20	Title	輕軌工業擬改採國內標	2	
	30	輕軌工業擬改採國內標，採國內外業者共同承攬但由國內業者主導。	1	G
	57	輕軌工業擬改採國內標，放寬招商「實績」要求，提高國內業者自製率比重至五０％，採國內外業者共同承攬但由國內業者主導。	1	G
21	Title	中共採購新規定，重擊微軟	2	
	36	中共採購新規定，重擊微軟，要購買非本國軟體系統的政府單位，一律特別呈報。	2	G
	63	中共採購新規定，重擊微軟，儘管微軟大力投資當地，並改組大中華區人事，但在大陸急力扶持國產軟件下，微軟在大陸業務可能遭致命打擊。	1	G
22	Title	扶持軟體產業，中共在融資、上市和稅收方面給予優惠措施	4	
	45	扶持軟體產業，中共在融資、上市和稅收方面給予優惠措施，成立風險投資公司，設立風險投資基金。	1	G
	62	扶持軟體產業，中共在融資、上市和稅收方面給予優惠措施，以求二〇一〇年大陸的軟體產業研究開發和生產能力達到或接近國際先進水平。	1	G
23	Title	緊縮房地產業，中共加大力道	3	
	40	緊縮房地產業，中共加大力道，要控制此類項目的建設用地供應量，或暫停審批此類項目。	1	G
	68	緊縮房地產業，中共加大力道，對高檔大戶型商品房、辦公大樓與商業性用房積壓較多的地區，要控制此類項目的建設用地供應量，或暫停審批此類項目。	1	G

24	Title	陳總統：中華民國是主權獨立國家	3	
	34	陳總統：中華民國是主權獨立國家，國軍要爲捍衛中華民國主權與領土而戰。	1	G
	66	外傳前總統李登輝指「陳總統只說中華民國是國號，沒有說中華民國是國家」，而陳總統昨天則向三軍官兵強調「中華民國是一個主權獨立的國家」。	3	G
25	Title	明年總統大選，藍綠基本盤皆見鬆動	3	
	45	明年總統大選，藍綠基本盤皆見鬆動，而當年的選民，歷經政黨輪替，如今投票意向已出現明顯改變。	1	G
	64	明年總統大選，藍綠基本盤皆見鬆動，上屆大選支持泛藍的選民，陣腳略微鬆動；而之前支持陳呂配的泛綠選民，也有相當比例出現流失的現象。	1	G
26	Title	競國實業董事會決議配息配股基準日爲 9 月 12 日。	1	
	39	競國實業董事會決議配息配股基準日爲 9 月 12 日，9 月 8 日起至 9 月 12 日停止股票過戶。	1	G
	39	競國實業董事會決議配息配股基準日爲 9 月 12 日，9 月 8 日起至 9 月 12 日停止股票過戶。	1	G
27	Title	國眾奪下中華電北區 FTTBL2Switch 採購案	2	
	39	國眾奪下中華電北區 FTTBL2Switch 採購案，以供中華電信協助中小企業利用寬頻網路發展商機之用。	1	G
	58	國眾奪下中華電北區 FTTBL2Switch 採購案，由國眾得標，智邦集團傳易（SMC）、和心光通、飛瑞、安捷倫及浩網等廠商負責提供相關整合產品。	1	G
28	Title	亞太電信集團跨足線上遊戲，今年營收約 2500 萬元	3	
	36	亞太電信集團跨足線上遊戲，今年營收約 2500 萬元，4C 整合的佈局儼然成形。	1	G
	69	亞太電信集團跨足線上遊戲，今年營收約 2500 萬元，推出新的娛樂事業群，亞太集團版圖橫跨了電信、網路、通訊、加值內容，4C 整合的佈局儼然成形。	1	G
29	Title	亞太電信推出「猿人在線」品牌，初期以代理爲主。	3	
	31	亞太電信推出「猿人在線」品牌，初期以代理爲主，朝線上遊戲邁進。	2	G
	53	亞太電信推出「猿人在線」品牌，初期以代理爲主，因此結合集團內各式寬頻服務載具與平台的資源，朝線上遊戲邁進。	1	F
30	Title	友達第五代彩色濾光片廠十月起逐步量產，最大月產能 12 萬片	3	
	43	友達第五代彩色濾光片廠十月起逐步量產，最大月產能 12 萬片，使友達有效掌握上游關鍵零組件。	1	G
	64	友達第五代彩色濾光片廠十月起逐步量產，最大月產能 12 萬片，月產能 7 萬片，預估未來每月最大產能 12 萬片玻璃基板，供全球大尺寸面板需求。	1	F
31	Title	中壽投資型商品「一觸得利」狂賣，一周銷售達 13 億元	5	
	40	中壽投資型商品「一觸得利」狂賣，一周銷售達 13 億元，不僅爲業界首創，引發熱賣風潮。	2	G
	57	中壽投資型商品「一觸得利」狂賣，一周銷售達 13 億元，投資標的爲逆浮動+正浮動利率債券，不僅爲業界首創，引發熱賣風潮。	2	G
32	Title	29 日台積電 ADR 收盤價 11.78 美元，較前交易日上漲 0.08 美元。	1	
	42	29 日台積電 ADR 收盤價 11.78 美元，較前交易日上漲 0.08 美元，漲幅爲 0.68%，換算回台股每股價格約 80.54 元。	1	G
	54	29 日台積電 ADR 收盤價 11.78 美元，較前交易日上漲 0.08 美元，較前一交易日上漲 0.08 美元，漲幅爲 0.68%，換算回台股每股價格約 80.54 元。	1	B

33	Title	「美夢成真」趕戲，葉全真累壞吊了點滴再上	3	
	45	「美夢成真」趕戲，葉全真累壞吊了點滴再上，不顧醫生要她吊點滴多休息的叮嚀，又回棚內拍戲去。	1	G
	59	「美夢成真」趕戲，葉全真累壞吊了點滴再上，所以她在打了兩劑粗血管針後，不顧醫生要她吊點滴多休息的叮嚀，又回棚內拍戲去。	1	B
34	Title	八點檔現拍現播，演員連連發病	4	
	43	八點檔現拍現播，演員連連發病，除了中視、華視，其餘三台都以現拍現播的方式，走本土路線。	1	G
	58	八點檔現拍現播，演員連連發病。演員日夜趕戲來趕播出，體力已受考驗，偏偏表演方式更耗費體力，病號、傷兵也因此連連爆發。	2	G
35	Title	周俊三蹲牢房，代價很值得	2	
	35	周俊三蹲牢房，代價很值得，辛苦還是有代價的，讓他獲得 3 萬元的豐厚酬勞。	1	G
	35	周俊三蹲牢房，代價很值得，辛苦還是有代價的，讓他獲得 3 萬元的豐厚酬勞。	1	G
36	Title	佼佼訪王貞治，豪華日本行。	1	
	45	佼佼訪王貞治，豪華日本行，還住在一晚高達 6 萬日幣的飯店裡，且如願吃到頂級的佐賀牛肉壽喜燒。	1	G
	67	佼佼訪王貞治，豪華日本行，除了能親眼目睹日本職棒，專訪職棒明星王貞治，還住在一晚高達 6 萬日幣的飯店裡，且如願吃到頂級的佐賀牛肉壽喜燒。	1	G
37	Title	「棋靈王圍棋入門之旅」活動開跑	3	
	34	「棋靈王圍棋入門之旅」活動開跑，使得圍棋儼然成爲最新的全民益智運動。	1	G
	61	「棋靈王圍棋入門之旅」活動開跑，再加上不久前奪得今年日本本因坊頭銜的旅日棋手張栩效應，使得圍棋儼然成爲最新的全民益智運動。	1	G
38	Title	周末官邸藝文沙龍，王琄邀親子無言的交流	5	
	35	周末官邸藝文沙龍，王琄邀親子無言的交流，激發出親子間的想像力與創造力！	1	G
	60	周末官邸藝文沙龍，王琄邀親子無言的交流，並且藉由各式精心設計的劇場遊戲一模仿、帶領、互動，激發出親子間的想像力與創造力！	1	G
39	Title	故宮德國文物大展，開放展場設計權	3	
	39	故宮德國文物大展，開放展場設計權，舉辦公開說明會，歡迎設計師與建築師前來參與。	2	G
	65	故宮德國文物大展，開放展場設計權，故宮破天荒將公開舉辦展場競圖，預計本月 15 日下午 2 點，舉辦公開說明會，歡迎設計師與建築師前來參與。	1	G
40	Title	藝文界前輩進駐爲豐樂童畫賽暖身	1	
	15	藝文界前輩進駐爲豐樂童畫賽暖身	1	B
	15	藝文界前輩進駐爲豐樂童畫賽暖身	1	B

Table 5 summarizes the data shown in Table 4. As can be seen, of the 40 stories, 65.0% or 62.5% of the first candidates suggested by the method for the length limit 45 and 69, respectively, were judged good. If users were able to choose from all the suggested candidates, 80% or 75% of the summaries could be obtained from a machine without manual editing. Only about 12.5% or 10% of the stories yielded summaries that were unacceptable.

Table 5. Quality statistics for the summary candidates of the inside testing set. (a) The upper table is for length limit 45. (b) The lower table is for length limit 69.

Quality / Rank	Good	Fair	Bad
1	26 (65.0%)	2 (5.0%)	5 (12.5%)
2	5 (12.5%)	1 (2.5%)	0
3	1 (2.5%)	0	0
4	0	0	0
5	0	0	0
total	32 (80.0%)	3 (7.5%)	5 (12.5%)

Quality / Rank	Good	Fair	Bad
1	25 (62.5%)	6 (15%)	4 (10%)
2	3 (7.5%)	0	0
3	1 (2.5%)	0	0
4	1 (2.5%)	0	0
5	0	0	0
total	30 (75.0%)	6 (15%)	4 (10%)

The best candidates that are unacceptable (9 cases in total in news ID 4, 5, 8, 15, 17, 32, 33, and 40) contain undesired conjunctions that break the coherence and/or readability (4 cases in 4, 5, 8, 33), clauses that duplicate the headline strings (2 cases in 17 and 32), or were nothing but the headline itself, which means that no candidates could be generated under the required length limit (3 cases in 15 and 40). The suitability of conjunctions for direct concatenation is difficult to judge, because some of them are helpful and some are not. The cases of headline duplication can be eliminated by duplication detection before concatenation. As to those candidates that contain only headlines, the clauses can be broken into smaller structures, such as phrases, for re-combination. This, however, would be a more difficult problem that would require more language analysis.

Table 6 summarizes the results for the outside testing set. As can be seen, the percentages of the first suggested candidates that were judged good decrease from 65% and 62.5% to 18.7% and 33.3%, respectively, for the length limit 45 and 69. The percentages that were judged good regardless of the rank position decreased from 80% to 53.3% for the length limit 45 and from 75% to 70.7% for the length limit 69, showing that the shorter the length limit, the less stable the method in performance. A large portion of the percentage moves to those

that were judged fair. This decrease in performance may due to the greater number of body sentences and the larger number of evaluators for the outside testing set. As more candidates (and evaluators) exist for selection, less coincidence exists for the same choice (and judgment) results. The only consistent result (compared to the inside testing set) is that those best candidates that were judged bad are still rare (less than 10%). This shows that the heuristic: "*concatenating the last clauses of the body sentence with the headline*" seems to work for Chinese news in this application.

Table 6. Quality statistics for the summary candidates of the outside testing set. (a) The upper table is for length limit 45. (b) The lower table is for length limit 69.

Quality / Rank	Good	Fair	Bad
1	14 (18.7%)	11 (14.7%)	2 (2.7%)
2	14 (18.7%)	8 (10.7%)	0 (0.00%)
3	8 (10.7%)	5 (6.7%)	2 (2.7%)
4	4 (5.3%)	1 (1.3%)	0 (0.00%)
5	0 (0.00%)	4 (5.3%)	2 (2.7%)
total	40 (53.3%)	29 (38.7%)	6 (8.00%)

Quality / Rank	Good	Fair	Bad
1	25 (33.3%)	5 (6.7%)	3 (4.0%)
2	13 (17.3%)	8 (10.7%)	1 (1.3%)
3	7 (9.3%)	3 (4.0%)	0 (0.0%)
4	4 (5.3%)	1 (1.3%)	0 (0.0%)
5	4 (5.3%)	1 (1.3%)	0 (0.0%)
total	53 (70.7%)	18 (24.0%)	4 (5.3%)

5. Discussion

The fact that the proposed method works for some stories is due to the characteristics of Chinese news. They tell stories in a successive sequence. Very few grammatical inversions within sentences and clauses are used. Chinese words have virtually no morphological variations. The clauses, especially at the rear part of a sentence, are sometimes quite independent of the front part. Headlines are given in a compact form to cover as many important facets as possible, such as who, what, where, when, why, and how. All of these characteristics make clause recombination a choice for summary generation. With this

heuristic strategy, the remaining work is to evaluate their fitness as summaries and rank them in a correct sense. For the news stories we tested, the proposed method applies to most of them with success. Nonetheless, for stories not of this type, such as editorials, commentaries, and lists of events, items, or prices, this method may fail. For the stories whose headlines are more eye-catching rather than informative, such that most content words do not appear in the headlines, this method may fail as well.

6. Conclusions

The proposed method recombines snippets of news without modifying them. A direct advantage is that other synchronized media such as images, speech, or video of the same story can maintain synchronization with ease when they are summarized as well (like those in (ANSES, n.d.)), because the positions of where to cut and paste are known during the generation of the summary candidates. Thus, to achieve speech or video segmentation and summarization for similar services, one can use their synchronized texts based on this method.

Other practical advantages of this computer-assisted summarization include the ease of maintaining summary quality regardless of the experience of human summarizers and the reduction in the cost and time to train novices for this kind of services.

Evaluation of the quality of auto-generated summaries requires human judgment and is, thus, expensive and time-consuming for large-scale or multiple-run evaluation. To allow automatic evaluation using the methodology like those used in machine translation (Papineni, Roukos, Ward, & Zhu, 2002; Doddington, 2002), a number of test collections need to be created. Our past research projects in Chinese OCR text retrieval and Chinese document classification have results in two corresponding test collections for free use (Tseng, 2002; 2004). We hope that we can also release a Chinese collection for evaluating automatic summarization in the future.

Acknowledgement

Special thanks are to the anonymous reviewers for their valuable comments and suggestions to improve the quality of this paper.

This work is partly supported by WebGenie Information Ltd. and National Science Council under the grants numbered: NSC 93-2213-E-030-007- and NSC 97-2631-S-003-003-.

References

ANSES : Automatic News Summarization and Extraction System : Knowledge Media Institute : The Open University. Available: http://technologies.kmi.open.ac.uk/anses/ [Accessed: July 22, 2009].

Barzilay, R. & McKeown, K. R. (2005). Sentence Fusion for Multidocument News Summarization. *Computational Linguistics*, 31(3), 297-327.

Carbonell, J. & Goldstein, J. (1998). The use of MMR, diversity-based re-ranking for reordering documents and producing summaries. In *Proceedings of SIGIR*, Melbourne, Australia, 335-336.

Central News Agency. Price List for Stock News Brief Message from Central News Agency. (in Chinese). Available: http://www.suio.com.tw/top/can/can_order_txt.asp [Accessed: July 22, 2009]

Chen, H.-H., Kuo, J.-J., Huang, S.-J., Lin, C.-J., & Wung, H.-C. (2003). A summarization system for Chinese news from multiple sources. *Journal of the American Society for Information Science and Technology*, 54(13), 1224-1236.

China Times. Media Challenges: Multi-modal communications. (in Chinese). Available: http://marketing.chinatimes.com/item_detail_page/professional_columnist/professional_columnist_content_by_author.asp?MMContentNoID=4369 [Accessed: Dec. 3, 2003].

Document Understanding Conferences, Available: http://duc.nist.gov/ [Accessed: July 22, 2009]

Doddington, G. (2002). Automatic Evaluation of Machine Translation Quality using N-gram Co-occurrence Statistics. In *Proceedings of the second international conference on Human Language Technology Research*, San Diego, California, March, 2002, 138-145.

Evans, D. K., Klavans, J. L., & McKeown, K. R. (2004). Columbia Newsblaster: Multilingual News Summarization on the Web. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, HLT-NAACL 2004: Demonstration Papers, May 2-7, 2004, Boston USA, 1-4.

Fukusima, T., Okumura, M., & Nanba, H. (2002). Text Summarization Challenge 2: Text Summarization Evaluation at NTCIR Workshop3. In *Proceedings of the Third NTCIR Workshop on Evaluation of Information Retrieval, Automatic Text Summarization and Question Answering*, Oct. 8-10, 2002, Tokyo, Japan, 1-6.

Hovy, E. H., & Lin, C.-Y. (1999). *Automated Text Summarization in SUMMARIST*. In I. Mani and M. Maybury, editors, Advances in Automated Text Summarization, chapter 8. MIT Press.

Levenshtein, V. I. (1966). Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady*, 10(8), 707-710.

Lin, C.-Y. & Hovy, E. H. (2002). From single to multi-document summarization: A prototype system and its evaluation. In *Proceedings of the ACL*, Philadelphia, PA, 2002, 457-464.

Lin, C.-Y. & Hovy, E. H. (2003). The Potential and Limitations of Sentence Extraction for Summarization. In *Proceedings of the Workshop on Automatic Summarization, post-conference workshop of HLT-NAACL-2003*, Edmonton, Canada, May 31 - June 1, 2003.

Lopresti, D. & Zhou, J. (1996). Retrieval Strategies for Noisy Text. In *Proceedings of the Fifth Annual Symposium on Document Analysis and Information Retrieval*, April 15-17, 1996, 255-269.

Inderjeet Mani, *Automatic Summarization*, John Benjamins, 2001.

Papineni, K., Roukos, S., Ward, T., & Zhu, W.-J. (2002). Bleu: A method for automatic evaluation of machine translation. In *Proceedings of the ACL*, Philadelphia, PA, 311-318.

Radev, D. R., Hovy, E., & McKeown, K. (2002). Introduction to the Special Issue on Summarization. *Computational Linguistics*, 28(4), 399-408.

Radev, D. R., Otterbacher, J., Winkel, A., & Blair-Goldensohn, S. (2005). NewsInEssence: summarizing online news topics. *Communications of the ACM*, 48(10), 95-98.

Takefumi, O., Hidetaka, M., & Hiroshi, N. (2003). Web News Article Summarization and its Evaluation using Articles for Mobile Terminals. *Joho Shori Gakkai Kenkyu Hokoku*, 2003(4), 1-8. (in Japanese)

Tseng, Y.-H. (2002). Automatic Thesaurus Generation for Chinese Documents. *Journal of the American Society for Information Science and Technology*, 53(13), 1130-1138.

Tseng, Y.-H. (2002). FJU Test Collection for Evaluation of Chinese OCR Text Retrieval. Available:http://www.lins.fju.edu.tw/~tseng/Collections/Chinese_OCR_IR.html [Accessed: July 22, 2009].

Tseng, Y.-H. (2004). FJU Test Collection for Evaluation of Chinese Text Categorization. Available: http://www.lins.fju.edu.tw/~tseng/Collections/Chinese_TC.html [Accessed: July 22, 2009].

Tseng, Y.-H., Wang, Y.-M., Lin, Y.-I., Lin, C.-J., & Juang, D.-W. (2007). Patent Surrogate Extraction and Evaluation in the Context of Patent Mapping. *Journal of Information Science*, 33(6), 718-736.

United Daily News. egolife: Brief Message Delivery for Cell Phone. (in Chinese) Available: http://udn.com/NASApp/LogFriend/UDNSMS/introduction_news.html [Accessed: Jan. 2, 2005]

Wang, C. (2002). Number One in the World: Cell Phone Numbers used More Than the Populations of Taiwan. (in Chinese) ETtoday.com, 2002/08/09. Available: http://www.ettoday.com/2002/08/09/339-1337800.htm [Accessed: July 22, 2009].

Computational Linguistics and Chinese Language Processing
Vol. 14, No. 1, March 2009, pp. 105-132

Study of Associative Cepstral Statistics Normalization Techniques for Robust Speech Recognition in Additive Noise Environments

Wen-Hsiang Tu* and Jeih-weih Hung*

Abstract

Feature statistics normalization techniques have been shown to be very successful in improving the noise robustness of a speech recognition system. In this paper, we propose an associative scheme in order to obtain a more accurate estimate of the statistical information in these techniques. By properly integrating codebook and utterance knowledge, the resulting associative cepstral mean subtraction (A-CMS), associative cepstral mean and variance normalization (A-CMVN), and associative histogram equalization (A-HEQ) behave significantly better than the conventional utterance-based and codebook-based versions in additive noise environments. For the Aurora-2 clean-condition training task, the new proposed associative histogram equalization (A-HEQ) provides an average recognition accuracy of 90.69%, which is better than utterance-based HEQ (87.67%) and codebook-based HEQ (86.00%).

Keywords: Speech Recognition, Noise-Robust Feature, Codebook

1. Introduction

The performance of a speech recognition system is often severely degraded when there is a mismatch between the acoustic conditions of the training and the application environments. This mismatch may come from various sources, such as additive noise, channel distortion, different speaker characteristics, and different speaking modes. A variety of robustness techniques with demonstrated improvement in system performance have been proposed to reduce this mismatch. For the purpose of handling additive noise, these robustness techniques can be roughly divided into three classes: adaptation of the speech models in the recognizer to make them better match the noise conditions, enhancement of the speech features before they are fed to the recognizer, and utilization of a noise robust representation of speech signals. In

* Dept of Electrical Engineering, National Chi Nan University, Nantou County, Taiwan, Republic of China
E-mail: aero3016@ms45.hinet.net; jwhung@ncnu.edu.tw

[Received January 13, 2009; Revised May 13, 2009; Accepted May 19, 2009]

the first class of approaches, compensation is performed on the pre-trained recognition model parameters so that the modified recognition models can more effectively classify the mismatched testing speech features collected in the application environment. Typical examples of this class include the well-known noise masking (Holmes & Sedgwick, 1986; Klatt, 1979; Nadas, Nahamoo, & Picheny, 1988), speech and noise decomposition (SND) (Varga & Moore, 1990), hypothesized Wiener filtering (Berstein & Shallom, 1991; Beattie & Young, 1992), vector Taylor series (VTS) (Acero, Deng, Kristjansson, & Zhang, 2000), maximum likelihood linear regression (MLLR) (Leggester & Woodland, 1995), model-based stochastic matching (Sankar & Lee, 1996; Lee, 1998), statistical re-estimation (STAR) (Moreno, Raj, & Stem, 1996), and parallel model combination (PMC) (Gales & Young, 1993; 1995a; 1995b). In the second class of approaches, the obtained testing speech features are modified in order to fit the acoustic conditions of pre-trained recognition models more compatibly. Examples of this class include the well-known spectral subtraction (SS) (Boll, 1979), codeword-dependent cepstral normalization (CDCN) (Acero, 1990), feature-based stochastic matching (Sankar & Lee, 1996; Lee, 1998), vector Taylor series (Segura, Benitez, de la Torre, Dupont, & Rubio, 2002; Moreno, Raj, & Stem, 1998), multivariate Gaussian-based cepstral normalization (RATZ) (Moreno, Raj, & Stem, 1996), and stereo-based piecewise linear compensation for environments (SPLICE) (Deng, Acero, Jiang, Droppo, & Huang, 2001; Droppo, Deng, & Acero, 2001). In the third class of approaches, a special robust speech feature representation is developed to reduce the sensitivity to various acoustic conditions; one way to develop this new feature representation is to normalize the statistics of the original speech features in both training and testing conditions in order to reduce the mismatch caused by noise. These feature statistics normalization techniques include cepstral mean subtraction (CMS) (Atal, 1974), cepstral mean and variance normalization (CMVN) (Tibrewala & Hermansky, 1997), cepstral gain normalization (CGN) (Yoshizawa, Hayasaka, Wada, & Miyanaga, 2004), histogram equalization (HEQ) (Hilger & Ney, 2006), higher-order cepstral moment normalization (HOCMN) (Hsu & Lee, 2004), cepstral shape normalization (CSN) (Du & Wang, 2008) *etc*. A common advantage of these methods is simplicity of implementation, since all of them focus on the front-end speech feature processing without the need of changing the back-end model training and recognition schemes. Regardless of the simplicity, these methods usually improve the recognition performance significantly under a noise-corrupted application environment.

A key process for most of the above normalization methods is to estimate the statistical information of speech features. For example, the first-order moment (mean), the first and second-order moments (mean and variance), and the probability distribution of features are required for CMS, CMVN, and HEQ, respectively. In most cases, the required statistical information is directly evaluated from the entire frame set of an utterance. Although simple in

implementation, the resulting utterance-based methods likely have some inherent drawbacks. First, they cannot be realized in an on-line manner since the computation and normalization of the statistics cannot be performed until the last frame of an utterance is received. Second, the number of frames in an utterance influences the accuracy of the obtained statistics. Third, since the length, or the number of different acoustic units, may vary from utterance to utterance, the normalized features of the same acoustic unit in an utterance may differ from those in another utterance.

In our previous works (Hung, 2006; 2008), we proposed that the statistics of features be evaluated based on two codebooks, named "pseudo stereo codebooks". Construction of the codebook of clean speech cepstra can occur off-line and prior to recognition. The codebook of noise-corrupted speech cepstra for each testing utterance is constructed by properly integrating the clean-speech codebook and the noise estimates, which often can be extracted from the first several frames of the utterance. The resulting codebook-based methods are expected to obtain more accurate estimate of feature statistics, and they can be implemented in an almost on-line manner. In (Hung, 2008), we have shown that codebook-based CMS and CMVN outperform conventional utterance-based ones in recognition accuracy for additive noise environments.

The original procedures in constructing the codebooks in (Hung, 2008), however, are somewhat simple, which possibly results in a less accurate estimate of the statistics for speech features. First of all, the clean speech codebook is built with all the feature vectors in the clean speech utterances for training. Since these utterances may contain quite long non-speech (silence) segments, it is likely that numerous codewords in the codebook just correspond to these non-speech parts. Second, the feature statistics are estimated by *uniformly* averaging the codewords, which ignores the relative significance of each codeword. Finally, the noise information only depends on the leading frames of an utterance, which may make the noise-corrupted speech codebook less accurate. This problem will be worse if the noise is non-stationary. Although updating the noise estimate within an utterance based on a voice activity detection (VAD) process can alleviate this problem, it will substantially increase the implementation complexity.

Based on the above observations, in this paper, we propose to improve the accuracy of the feature statistics estimation in two aspects. First, the procedures of creating the pseudo stereo codebooks are modified so that they are more representative of the speech features. The resulting advanced pseudo stereo codebooks are shown to be more effective in the codebook-based methods than the original ones. Second, the information from both the codebook and the frames of the processed utterance are integrated, so that more accurate statistics of the features can be obtained in order to further enhance the feature statistics normalization techniques. This idea is realized on three well-known approaches, CMS, CMVN, and HEQ. We will show that the resulting "associative" methods are superior to the original

utterance-based and codebook-based ones in the Aurora-2 clean-condition training task.

The remainder of the paper is organized as follows: Section 2 presents the construction of advanced pseudo stereo codebooks. Section 3 introduces our proposed associative cepstral normalization techniques. The experimental environment setup is described in Section 4, and the recognition results are given and discussed in Section 5. Finally, Section 6 contains brief conclusions.

2. The Construction of Advanced Pseudo Stereo Codebooks

In this section, we introduce the approach to constructing the advanced pseudo stereo codebooks for clean training and noise-corrupted testing environments, respectively. The corresponding procedures are also shown in Figure 1. The basic idea of the process for constructing these codebooks is as follows: during the feature extraction processes, we find an intermediate feature domain in which the clean speech and noise are *linearly additive* (assuming that the speech signal and noise are uncorrelated in the time domain). The clean speech codewords for the intermediate feature domain first are constructed then are linearly added to the noise estimates to compose the noise-corrupted speech codewords for that domain. Finally, they are transformed to the final feature domain following the remaining feature extraction processes. For the mel-frequency cepstral coefficients (MFCC), the intermediate feature mentioned above is the mel-spectrum, while for the other two types of speech features, linear prediction cepstral coefficients (LPCC) (Atal, 1974; Makhoul, 1975) and perceptual linear prediction cepstral coefficients (PLPCC) (Hermansky, 1990), both the auto-correlation coefficients and the magnitude spectrum can be selected as the intermediate feature. Therefore, the codebook construction process and the relating methods can be applied to MFCC, LPCC, and PLPCC.

For simplicity, the mel-frequency cepstral coefficients (MFCC) are used as the speech features here; thus, the two codebooks are just designed for MFCCs. Following the derivation processes of the MFCCs for a speech signal, the speech portions of all clean speech utterances in the training database are converted into sequences of mel-spectral vectors, each consisting of the mel-filter bank outputs. These vectors are then used to construct a set of R codewords together with their weights by vector quantization (VQ), denoted as:

$$\left\{\tilde{\mathbf{x}}[r], w_r;\ 1 \le r \le R\right\}, \tag{1}$$

where $\tilde{\mathbf{x}}[r]$ and w_r represent the r^{th} codeword and its corresponding weight, respectively. Each weight w_r represents the relative cluster size in VQ classification. These mel-spectral codewords are then transformed into the cepstral domain as follows:

$$\mathbf{x}[r] = \mathbf{C}\log\left(\tilde{\mathbf{x}}[r]\right), \tag{2}$$

where $\mathbf{C}$ is the discrete-cosine-transform (DCT) matrix. Thus, the set of codewords $\left\{\tilde{\mathbf{x}}[r], w_r;\ 1 \le r \le R\right\}$ is the clean speech cepstral codebook. Note that we construct this cepstral codebook by transforming the mel-spectral codewords rather than by vector quantizing the cepstral features directly and that these mel-spectral codewords are preserved in order to construct the noise-corrupted speech codebook.

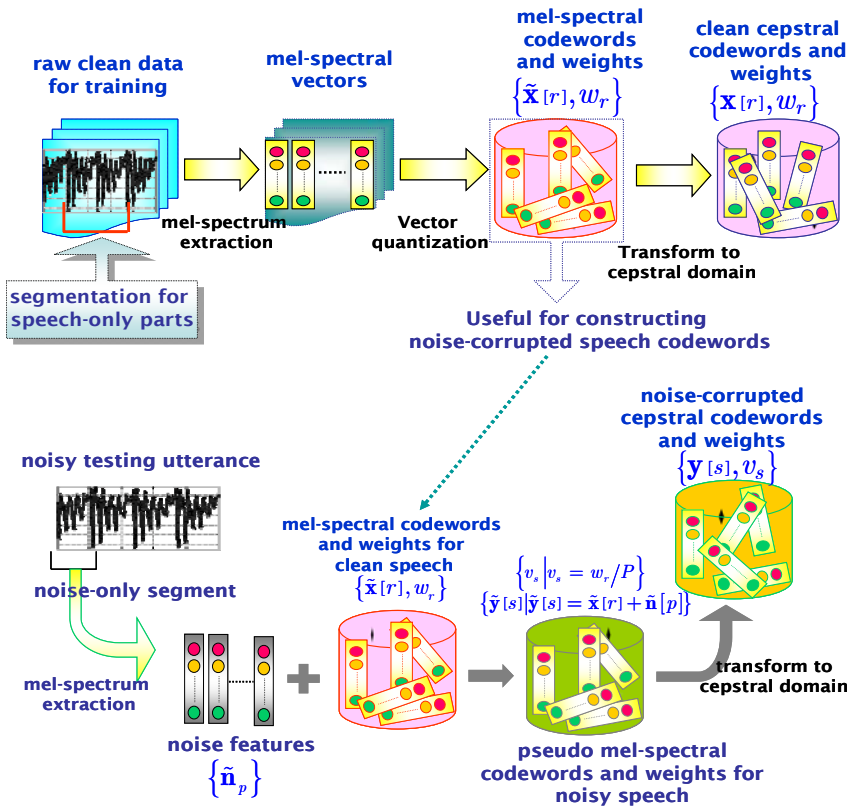

Figure 1. The procedures of constructing advanced pseudo stereo codebooks.

For the noise-corrupted testing environment, since it is often difficult to obtain a set of reliable codewords completely based on a single testing utterance, we construct the "noise-corrupted speech" codebook with the help of the available "clean-speech" mel-spectral codewords. For a given noise-corrupted testing utterance, let the estimated mel-spectra of the noise be approximated as a set of vectors, which are denoted as $\left\{\tilde{\mathbf{n}}[p];1 \le p \le P\right\}$, where P is the number of noise vectors. Then, since the clean speech and noise are approximately additive in the mel-spectral domain, the mel-spectral codewords for the noise-corrupted utterance are obtained as

$$\tilde{\mathbf{y}}[s]\Big|_{s=(r-1)P+p} = \tilde{\mathbf{x}}[r] + \tilde{\mathbf{n}}[p]. \tag{3}$$

and the weight for each $\tilde{\mathbf{y}}[s]$ is approximated by

$$v_s\Big|_{s=(r-1)P+p} = w_r / P\,. \tag{4}$$

Finally, we transform each $\tilde{\mathbf{y}}[s]$ into the cepstral domain, as in Eq. (2):

$$\mathbf{y}[s] = \mathbf{C}\log\left(\tilde{\mathbf{y}}[s]\right), \tag{5}$$

where $\mathbf{C}$ is the discrete-cosine-transform (DCT) matrix. Thus, $\left\{\mathbf{y}[s], v_s; 1 \le s \le RP\right\}$ is the noise-corrupted speech cepstral codebook. From the above, the two sets of codewords, $\left\{\mathbf{x}[r], w_r\right\}$ and $\left\{\mathbf{y}[s], v_s\right\}$, are viewed as the representatives for the clean training and noise-corrupted testing conditions, respectively, and they are named "pseudo stereo codebooks" here. The term "pseudo" indicates that the noise-corrupted speech codebook is *not* derived from the noise-corrupted speech directly, but is a fusion of the clean speech codebook and the noise estimates.

Here, the codebook construction scheme is different from that in (Hilger & Ney, 2006) in two points:

1. A speech/non-speech classification, or voice activity detection (VAD) procedure, is performed on each utterance in the clean training database, and then *only the speech portions* of these utterances are used to construct the clean-speech codewords. Thus, the resulting codewords will convey the speech characteristics better than those obtained in (Hilger & Ney, 2006). More precisely, for almost every utterance in the clean training set, there is a relatively long silent portion preceding and/or following the speech-containing portion. Therefore, if we create the codewords with utterances that are not VAD-processed, there may be a significant number of codewords that correspond to the silence, or the codewords corresponding to the silence may possess relatively high weight values, which will result in a less accurate estimate for the statistics of speech features. Furthermore, since the utterances to be VAD-processed here are clean from noise, the results of speech/non-speech classification are very accurate.
2. Second, the codeword weights, $\left\{w_r\right\}$ and $\left\{v_s\right\}$, are additionally calculated. They represent the relative significance of each codeword; thus, the estimated statistics of the speech features based on the advanced codebooks are expected to be more accurate than those on the original ones in (Hung, 2008).

In the above codebook construction process, we only focus on the speech characteristics in an utterance for clean and noise-corrupted conditions, while the non-speech portions

(silence or noise-only regions) are not considered. In other words, a "non-speech" codebook is not constructed here. The main reason is simple: in the feature statistics normalization approaches, we often do not process the speech and non-speech frames separately in an utterance, but treat them in the same manner as the same estimated feature statistics. Although normalizing the speech and non-speech frames based on different feature statistics may bring better recognition performance, it requires a reliable voice activity detector (VAD) which classifies a frame as speech or non-speech accurately in both clean and noise-corrupted conditions. Nevertheless, in general, a VAD brings about higher mis-classification rates as the signal-to-noise ratio (SNR) gets worse in the noise-corrupted condition, thus possibly harming the performance of feature statistics normalization approaches.

With the help of the above advanced pseudo stereo codebooks, we can estimate the statistics and probability distribution functions for the features of both the clean and noise-corrupted speech. For example, given the time stream of the m^{th} cepstral coefficients, $\{c_m[n]\}$, of *a clean utterance in the training set*, the k^{th} order moment and the probability distribution function of C_m are approximated by

$$E\left\{\left(C_m\right)^k\right\} \approx \sum_{r=1}^{R} w_r \left(x_m[r]\right)^k, \quad k = 1,2,3,\cdots, \tag{6}$$

and

$$F_{C_m}\left(z\right) \triangleq P\left(C_m \le z\right) \approx \sum_{r=1}^{R} w_r u\left(z - x_m[r]\right), \tag{7}$$

respectively, where C_m denotes a random variable with the samples $\{c_m[n]\}$, $x_m[r]$ is the m^{th} component of the clean speech codeword $\mathbf{x}[r]$, and $u(.)$ is the unit step function, defined by

$$u\left(\ell\right) = \begin{cases} 1 & \text{if } \ell \ge 0 \\ 0 & \text{if } \ell < 0 \end{cases}. \tag{8}$$

Similarly, if the time stream $\{c_m[n]\}$ corresponds to *a noise-corrupted utterance in the testing set*, then the k^{th} order moment and the probability distribution function of C_m are approximated by:

$$E\left\{\left(C_m\right)^k\right\} \approx \sum_{s=1}^{RP} v_s \left(y_m[s]\right)^k, \quad k = 1,2,3,\cdots, \tag{9}$$

and

$$F_{C_m}\left(z\right) \triangleq P\left(C_m \le z\right) \approx \sum_{s=1}^{RP} v_s u\left(z - y_m[s]\right) \tag{10}$$

respectively, where $y_m[s]$ is the m^{th} component of the noise-corrupted speech codeword $\mathbf{y}[s]$.

Based on these estimated statistics from the advanced codebooks, as in Eqs. (6), (7), (9), and (10), the codebook-based cepstral statistics normalization methods stated in (Hung, 2008) can be implemented. In Section 5, we will show that the advanced codebooks benefit the codebook-based CMS and CMVN in improving the recognition accuracy under noise-corrupted environments.

3. Associative Cepstral Statistics Normalization Techniques

The previous section introduces how to construct a better codebook set in order to enhance the corresponding codebook-based feature normalization methods. Updating the noise information in the noise-corrupted codebook, however, especially for a non-stationary noise environment, still makes the normalization method less efficient in computation. Besides, the codebook-based methods do not behave very well for some normalization methods, like HEQ, which will be shown in the subsequent sections. As a result, we attempt to incorporate the whole-utterance frames with the developed codebooks to evaluate the feature statistics, in the hope that the resulting feature statistics normalization techniques can bring better recognition accuracy. We realize our idea in the methods of CMS, CMVN, and HEQ, respectively, which is described in the following three subsections.

3.1 Associative Cepstral Mean Subtraction

Cepstral mean subtraction (CMS) (Atal, 1974) is a well-known speech feature processing technique. It was initially developed for eliminating the channel distortion in the features, but was found to be helpful as well in alleviating the effect of additive noise. In CMS, the original features are normalized to have zero mean. Briefly speaking, with the time-trajectory of the m^{th} cepstral coefficients, $\{c_m[n]\}$, for an utterance as the input, the output of the CMS process is expressed as:

$$\tilde{c}_m[n] = c_m[n] - \mu_m, \quad 1 \le n \le N, \tag{11}$$

where N is the number of frames in the utterance, and μ_m is the mean (the first-order moment) of $c_m[n]$. Here, the parameter μ_m is estimated by incorporating the codebooks, $\{\mathbf{x}[r], w_r\}$ and $\{\mathbf{y}[s], v_s\}$, in Section 2 and the whole-utterance frames, $\{c_m[n], 1 \le n \le N\}$. That is, for a clean speech utterance in the training set,

$$\mu_m = \alpha\left(\sum_{r=1}^{R} w_r x_m[r]\right) + (1-\alpha)\left(\frac{1}{N}\sum_{n=1}^{N} c_m[n]\right), \tag{12}$$

and for a noise-corrupted speech utterance in the testing set,

$$\mu_m = \alpha\left(\sum_{s=1}^{RP} v_s y_m[s]\right) + (1-\alpha)\left(\frac{1}{N}\sum_{n=1}^{N} c_m[n]\right). \tag{13}$$

In Equations (12) and (13), $x_m[r]$ and $y_m[s]$ denote the m^{th} component of the codewords $\mathbf{x}[r]$ and $\mathbf{y}[s]$, respectively, and α is a weighting factor between 0 and 1, which determines the usage ratio between the codebook and the whole-utterance frames. Here, the CMS method with the mean parameters defined in Eqs. (12) and (13) is named associative CMS (A-CMS). Obviously, if α is set to 1, the information from the frames in the utterance is completely ignored, and A-CMS is identical to codebook-based CMS (C-CMS). On the other hand, A-CMS with $\alpha = 0$ behaves equally to utterance-based CMS (U-CMS).

3.2 Associative Cepstral Mean and Variance Normalization

In the method of cepstral mean and variance normalization (CMVN) (Tibrewala & Hemansky, 1997), the original features are normalized to have zero mean and unity variance. With the time-trajectory of the m^{th} cepstral coefficients, $\{c_m[n]\}$, for an utterance as the input, the output of the CMVN process is expressed as:

$$\tilde{c}_m[n] = \left(c_m[n] - \mu_m\right)\Big/\sigma_m, \quad 1 \le n \le N, \tag{14}$$

where N is the number of frames in the utterance, while μ_m and σ_m are the mean and standard deviation of $c_m[n]$, respectively. In general, CMVN performs better than CMS because it additionally normalizes the variance of the features.

Similar to the previous sub-section, we estimate the two parameters, μ_m and σ_m, by incorporating the codebooks and the whole-utterance frames. That is, for a clean speech utterance in the training set,

$$\mu_m = \alpha\left(\sum_{r=1}^{R} w_r x_m[r]\right) + (1-\alpha)\left(\frac{1}{N}\sum_{n=1}^{N} c_m[n]\right), \tag{15}$$

$$\sigma_m^2 = \alpha\left(\sum_{r=1}^{R} w_r x_m^2[r]\right) + (1-\alpha)\left(\frac{1}{N}\sum_{n=1}^{N} c_m^2[n]\right) - \mu_m^2, \tag{16}$$

and for a noise-corrupted speech utterance in the testing set,

$$\mu_m = \alpha\left(\sum_{s=1}^{RP} v_s y_m[s]\right) + (1-\alpha)\left(\frac{1}{N}\sum_{n=1}^{N} c_m[n]\right), \tag{17}$$

$$\sigma_m^2 = \alpha\left(\sum_{s=1}^{RP} v_s y_m^2[s]\right) + (1-\alpha)\left(\frac{1}{N}\sum_{n=1}^{N} c_m^2[n]\right) - \mu_m^2. \tag{18}$$

In Equations (15)-(18), $x_m[r]$ and $y_m[s]$ denote the m^{th} component of the codewords $\mathbf{x}[r]$ and $\mathbf{y}[s]$, respectively, and α is a weighting factor between 0 and 1, which

determines the usage ratio between the codebook and the whole-utterance frames. Here, the CMVN method with means and variances defined in Eqs. (15)-(18) is named associative CMVN (A-CMVN). Similar to the case in the previous subsection, A-CMVN with $\alpha = 1$ is equivalent to codebook-based CMVN (C-CMVN), while A-CMVN with $\alpha = 0$ behaves equally to utterance-based CMVN (U-CMVN).

3.3 Associative Histogram Equalization

The histogram equalization (HEQ) technique (Hsu & Lee, 2004) normalizes each cepstral component stream so that the resulting histogram is close to a reference function. Following the notation of the previous two subsections, with the time-trajectory of the m^{th} cepstral coefficients $\left\{c_m[n]\right\}$ for an utterance as the input, the output of the HEQ process can be expressed as:

$$\tilde{c}_m[n] = F_N^{-1}\left(F_{C_m}\left(c_m[n]\right)\right), \quad 1 \le n \le N, \tag{19}$$

where $F_{C_m}(.)$ is the probability distribution function of $\left\{c_m[n]\right\}$, and $F_N(.)$ is a pre-defined reference distribution function. Compared with CMS and CMVN, HEQ additionally compensates all the higher-order moments of the features, and this extra compensation often results in an apparent improvement.

Analogous to the previous subsections, the distribution function $F_{C_m}(.)$ is jointly determined by the codebooks and the whole-utterance frames, and the resulting algorithm is called associative HEQ (A-HEQ). In A-HEQ, for a clean speech utterance in the training set,

$$F_{C_m}(z) = \alpha\left(\sum_{r=1}^{R} w_r u\left(z - x_m[r]\right)\right) + (1-\alpha)\left(\frac{1}{N}\sum_{n=1}^{N} u\left(z - c_m[n]\right)\right), \tag{20}$$

and for a noise-corrupted speech utterance in the testing set,

$$F_{C_m}(z) = \alpha\left(\sum_{s=1}^{RP} v_s u\left(z - y_m[s]\right)\right) + (1-\alpha)\left(\frac{1}{N}\sum_{n=1}^{N} u\left(z - c_m[n]\right)\right), \tag{21}$$

where $u(.)$ is the unit step function, as in Eq. (8).

Again, in Equations (20) and (21), the weighting factor α determines the usage ratio between the codebook and the whole-utterance frames. In the extreme case, $\alpha = 1$, the distribution function is completely determined by the codebook, thus A-HEQ becomes codebook-based HEQ (C-HEQ). In the other extreme case of $\alpha = 0$, A-HEQ corresponds to utterance-based HEQ (U-HEQ).

3.4 Comparison with Some Other Noise Compensation Algorithms

Previous work presents a series of noise compensation approaches which consider the speech and noise characteristics simultaneously, including the parallel model combination (PMC) (Gales & Young, 1993; 1995a; 1995b), vector Taylor series (VTS) (Acero, Deng, Kristjansson, & Zhang, 2000; Segura, Benitez, de la Torre, Dupont, & Rubio, 2002; Moreno, Raj, & Stem, 1998), and stereo-based piecewise linear compensation for environments (SPLICE) (Deng, Acero, Jiang, Droppo, & Huang, 2001; Droppo, Deng, & Acero, 2001). Here, we discuss the relationship of our proposed methods with PMC, VTS, and SPLICE, as well as the differences among them as follows:

1. In PMC, the original clean speech model parameters in the cepstral domain are transformed to the linear spectral domain, combined with the noise model parameters, then transformed back to the cepstral domain to be the approximated noisy speech model. Therefore, PMC compensates the speech model while keeping the noisy testing speech unchanged. Similar to PMC, in our proposed methods, the noisy speech codewords are obtained by integrating the clean speech codewords and the noise estimates in the linear spectral domain. Nevertheless, in our method, both the clean training and noisy testing speech data are compensated, then the speech model is trained (not just modified) with the new clean training speech data.
2. The VTS algorithm is often applied in two directions: one to compensate the speech model while keep the noisy testing speech unchanged, and the other to compensate the noisy testing speech without altering the original clean speech model. Briefly speaking, VTS considers that noisy speech is a nonlinear function of clean speech and noise in the logarithmic spectral domain and that this nonlinear function is approximated as a polynomial in order to estimate the statistics of noisy speech with the statistics of clean speech of noise. Therefore, our proposed methods differ from VTS in two ways: both the noisy testing speech and the speech model are changed in our methods and we primarily deal with the relationship of clean speech and noise in the linear spectral domain.
3. In SPLICE, the restored clean speech cepstral vector is obtained by adding the noisy speech cepstral vector to a correction vector. The correction vector is trained using the stereo recordings for both the clean and noisy speech data based on the maximum likelihood principle. In fact, in SPLICE, a Gaussian mixture model (GMM) for noisy speech cepstral vectors is trained, and the minimum-mean-square-error (MMSE) rule or the approximate maximum *a posteriori* (MAP) rule is applied to obtain the optimal estimate of the clean speech cepstral vector, given the noisy speech cepstral vector. Therefore, compared with SPLICE, our proposed methods do not use the stereo data since the noisy speech codebook is constructed simply by integrating the clean speech

codewords and the noise estimates. In addition, in SPLICE, the VQ process is performed on the noisy speech data in the cepstral domain, while in our methods we implement the VQ process on the clean speech data in the linear spectral domain.

To sum up briefly, in PMC, VTS, and SPLICE, the clean speech or noisy speech is often modeled by a single Gaussian or a mixture of Gaussians, while our proposed methods the speech are partially represented by a set of codewords. Furthermore, our proposed methods have lower computation complexity than PMC, VTS, and SPLICE, while they can provide very good recognition performance, as will be shown in the next section.

4. Experimental Setup

The proposed codebook-based algorithms have been tested with the AURORA-Project Digit Database Version 2.0, which is described in detail in (Hirsch & Pearce, 2000). In this database, the recordings have been manually segmented into utterances, and each utterance is saved as a file. The number of digits in an utterance can be one, two, three, four, five, six, and seven. Besides the digits, there is always a silent section at the beginning and end of an utterance. The length of an utterance may vary from 0.59 sec to 5.15 sec, depending on the number of digits in the utterance. The testing data consist of 4004 utterances from 52 male and 52 female speakers. Three different subsets are defined: Test Set A and Test Set B are each affected by four types of noise, and Test Set C is affected by two types. The noises included are: subway, babble, car, exhibition, restaurant, street, airport, and train station. Each noise is added to the clean speech under seven different signal-to-noise ratios (SNRs): -5dB to 20dB, spaced in 5dB intervals, and clean (no noise). The signals in Set A and Set B are filtered with a G.712 filter, and those in Set C are filtered with a MIRS filter. G.712 and MIRS are two standard frequency characteristics defined by the ITU (ITU recommendation G.712, 1996). Since the proposed methods are focused on improving the recognition accuracy for an additive noise environment, only Set A and Set B are used for the subsequent experiments.

On the other hand, under the clean training condition, the training data consist of 8440 clean speech utterances produced by 55 male and 55 female adults. These signals are filtered with a G.712 filter without noise added. For the clean training phase, the 8440 strings in the training set are first processed by an energy-based VAD process (Tai & Hung, 2006), and the speech portions are converted into vector streams of 23 mel-spectral coefficients. All of the 23-dimensional feature vectors are used to construct a set of R codewords via vector quantization (VQ) with the *K*-means clustering algorithm, in which the squared Euclidean distance is used for VQ classification. These codewords are also converted to 13-dimensional mel-frequency cepstral vectors (c_0~c_{12}) to form the clean speech cepstral codebook $\left\{\mathbf{x}[r], w_r; 1 \le r \le R\right\}$. Besides, all 8440 strings (including speech and non-speech portions) in the training set are converted to MFCC feature vector streams. The resulting 13-dimensional

cesptral features plus their delta and delta-delta comprise the components of the final 39-dimensional feature vectors. With these feature vectors in the training set, two sets of hidden Markov models (HMMs) for each digit (oh, zero, one, …, eight, and nine) and silence are trained. The first set follows the Microsoft complex back-end training scripts (Droppo, Deng, & Acero, 2002), in which each digit HMM has 16 states and 20 Gaussian mixtures per state. The second set follows the standard training scripts provided in the Aurora-2 database (Hirsch & Pearce, 2000), in which each digit HMM has 16 states and 3 Gaussian mixtures per state.

For the testing phase, the leading 10 frames (0.1 sec) of each utterance are assumed to be noise-only, and their corresponding 23-dimensional mel-spectral vectors are the elements of the estimated noise components $\left\{\tilde{\mathbf{n}}[p];\ 1 \le p \le P\right\}$ (where $P = 10$). We then construct the noise-corrupted speech cepstral codebook $\left\{\mathbf{y}[s], v_s; 1 \le s \le RP\right\}$ following the procedures in Section 2. Based on the two codebooks $\left\{\mathbf{x}[r], w_r\right\}$ and $\left\{\mathbf{y}[s], v_s\right\}$, the proposed algorithms are performed to adjust the features for both training and testing. The reference distribution function in Equation (19) for HEQ is a Gaussian distribution with zero mean and unity variance. Note that, even though it has been shown that dynamically determining the noise-only components within an utterance based on a voice activity detector (VAD) improves the recognition performance of codebook-based methods (Hung, 2008), it will significantly increase the computation complexity, especially when the detected non-speech portion is quite long (the size of noise-corrupted speech codebook, RP, is proportional to the number of noise components, P). Besides, the results of VAD become less reliable when the signal-to-noise ratio (SNR) gets worse, which somewhat deteriorates the accuracy of the resulting noise-corrupted speech codebook. Based on the above observations, we select the first 0.1 second (10 frames) of each utterance as the noise-only components, in which there is always no speech.

5. Experimental Results and Discussions

We compare and analyze the recognition accuracy achieved by the different approaches proposed here for the AURORA-2 experimental environment. This includes 5 subsections. In Subsections 5.1-5.4, the experimental results are obtained via the first set of HMMs (the complex back-end). Subsection 5.5, on the other hand, presents the experimental results obtained via the second set of HMMs (the standard back-end). The results in the first four subsections help us investigate the best possible recognition accuracy achieved by the proposed methods with a more elaborate model structure, while the results in the last subsection can be used with the purpose of performance comparison with many other robustness techniques which are evaluated under the standard model structure.

Briefly speaking, in Subsection 5.1 we examine the advanced pseudo stereo codebooks

proposed in Section 2 to see if they bring better recognition results in codebook-based CMS and CMVN than the original ones proposed in (Hung, 2008). In Subsection 5.2, the proposed associative CMS, CMVN, and HEQ in Section 3 are compared with the corresponding utterance-based and codebook-based ones in terms of the recognition performance. In Subsection 5.3, the effect of the weighting factor α in associative approaches of Section 3 is analyzed to see the corresponding influence on the recognition performance. In Subsection 5.4, we compare the proposed approaches with some other noise-robust techniques. Finally, the proposed methods are evaluated with the standard back-end in Subsection 5.5.

5.1 Comparison of the Advanced Pseudo Stereo Codebooks with the Original Ones in Codebook-based Approaches

We compare the new proposed pseudo stereo codebooks in Section 2 with the original ones in (Hung, 2008) in terms of the recognition accuracy for the codebook-based CMS and CMVN. For the pseudo stereo codebooks, the number of clean speech codewords, R, is set to 16, 64, and 256, and the first 10 frames of each utterance are assumed to be noise-only and their corresponding features constitute the estimated noise vectors $\left\{\tilde{\mathbf{n}}[p],\ 1 \le p \le P = 10\right\}$. Thus, the number of noise-corrupted speech codewords, RP, is equal to 160, 640, or 2560.

Tables 1 and 2 list the recognition results of codebook-based CMS and CMVN, respectively, where these results present individual set recognition accuracy rates averaged over five SNR conditions (0~20dB, at 5dB intervals). For the purpose of clarification in the tables, a superscript "$^{(o)}$" is added to the names of C-CMS and C-CMVN to indicate that the two methods are based on the original pseudo stereo codebooks.

From the two tables, several phenomena can be found:

1. Both CMS and CMVN bring improvement in recognition accuracy when compared with the baseline processing. As expected, however, CMVN performs better than CMS in all cases. As a result, performing the variance normalization really helps improve the noise robustness of speech features.

2. Under the same assignment for the parameter R (the number of clean speech codewords), the advanced pseudo stereo codebooks provide C-CMS and C-CMVN with significantly better recognition accuracy than the original codebooks do. These results support our statement in Section 2 that the proposed new codebooks are capable of providing a better estimate of feature statistics, thus, benefit the codebook-based approaches.

3. In most cases, increasing the number of clean speech codewords R brings about improved recognition accuracy for the four methods, C-CMS$^{(o)}$, C-CMS, C-CMVN$^{(o)}$, and C-CMVN. For both C-CMS and C-CMVN with the advanced codebooks, however, a

moderate number of codewords already give rise to nearly optimal performance. Nevertheless, this is not the case for C-CMS$^{(o)}$ and C-CMVN$^{(o)}$, with the possible reason that, in these two methods, we have to increase the number of codewords so that more of them can represent the speech portions, and more accurate feature statistics can be estimated accordingly. As a result, it shows that the advanced codebooks with a moderate size can serve as good representatives for speech features.

Table 1. Recognition accuracy (%) achieved by two versions of codebook-based CMS with a different number of clean speech codewords R averaged across the SNRs between 0 and 20dB, under the complex back-end structure. AR (%) and RR (%) are the absolute and relative error rate reductions over the baseline. C-CMS uses the advanced pseudo stereo codebooks, while C-CMS$^{(o)}$ uses the original pseudo stereo codebooks in (Hung, 2008).

Method	Set A	Set B	Average	AR	RR
Baseline	71.92	67.79	69.86	—	—
C-CMS$^{(o)}$ ($R = 16$)	74.21	70.81	72.51	2.65	8.81
C-CMS ($R = 16$)	79.04	79.56	79.30	9.45	31.33
C-CMS$^{(o)}$ ($R = 64$)	74.03	70.74	72.39	2.53	8.39
C-CMS ($R = 64$)	80.79	80.19	80.49	10.64	35.28
C-CMS$^{(o)}$ ($R = 256$)	77.92	75.20	76.56	6.71	22.24
C-CMS ($R = 256$)	81.46	81.49	81.48	11.62	38.55

Table 2. Recognition accuracy (%) achieved by two versions of codebook-based CMVN with a different number of clean speech codewords R averaged across the SNRs between 0 and 20dB, under the complex back-end structure. AR (%) and RR (%) are the absolute and relative error rate reductions over the baseline. C-CMVN uses the advanced pseudo stereo codebooks, while C-CMVN$^{(o)}$ uses the original pseudo stereo codebooks in (Hung, 2008).

Method	Set A	Set B	average	AR	RR
Baseline	71.92	67.79	69.86	—	—
C-CMVN$^{(o)}$ ($R = 16$)	84.44	82.40	83.42	13.57	45.00
C-CMVN ($R = 16$)	85.41	85.21	85.31	15.46	51.27
C-CMVN$^{(o)}$ ($R = 64$)	84.13	81.53	82.83	12.98	43.04
C-CMVN ($R = 64$)	86.92	86.81	86.87	17.01	56.43
C-CMVN$^{(o)}$ ($R = 256$)	86.67	86.25	86.46	16.61	55.08
C-CMVN ($R = 256$)	87.10	87.32	87.21	17.36	57.57

5.2 Comparison of the Associative CMS, CMVN and HEQ with the Utterance-based and Codebook-based Approaches

The proposed associative cepstral normalization methods (A-CMS, A-CMVN, and A-HEQ) are evaluated here in terms of their robustness against noise. For the purpose of comparison, the experiments for the corresponding utterance-based and codebook-based methods are also performed. Here, the weighting factor α in Equations (12), (13), (15)-(18), (20), and (21) is preliminarily set to 0.5. Similar to the previous subsection, the size of the clean speech codebook, *R*, is set to 16, 64, or 256, and the number of leading frames for noise estimation, *P*, is set to 10.

Tables 3, 4, and 5 list the recognition results of various types of CMS, CMVN, and HEQ, respectively, where these results present individual set recognition accuracy rates averaged over five SNR conditions (0~20dB, at 5dB intervals). For example, in Table 3, the recognition accuracy rates for utterance-based CMS (U-CMS), codebook-based CMS (C-CMS) with $R = 256$, and associative CMS (A-CMS) with three assignments of the parameter R, are presented. A similar arrangement holds for Tables 4 and 5. From the three tables, a series of observations are obtained as follows:

1. Among the three types of CMS, A-CMS performs the best, followed by C-CMS and then U-CMS. This condition also holds for A-CMVN, C-CMVN, and U-CMVN. First, the results agree with those obtained in (Hung, 2008) that codebook-based CMS and CMVN behave better than utterance-based ones. Second, the associative CMS (A-CMS) and CMVN (A-CMVN) always outperform both their corresponding utterance-based and codebook-based ones. Therefore, combining codebooks with the processed utterance features in estimating the feature statistics indeed helps improve the recognition accuracy considerably.

2. For the three utterance-based methods, HEQ always performs better than CMVN and CMS. As stated in Section 3, compared with CMVN and CMS, HEQ additionally compensates for all the higher-order moments of features, thus, brings about extra improvement. This, however, is not the case for codebook-based methods: C-HEQ behaves worse than C-CMVN and is the worst of the three HEQ methods. A possible reason is that the codebooks give more accurate gross information (*i.e*. the mean and variance) for the features, but they are less capable of providing the detailed behavior (*i.e*. the probability distribution) of them.

3. Similar to the case in CMS and CMVN, the new A-HEQ outperforms U-HEQ and C-HEQ significantly. The superior performance of A-HEQ again reveals that the feature statistics can be estimated more accurately by incorporating the codebook and the processed utterance features.

4. In these high-performance A-CMS, A-CMVN, and A-HEQ, setting the weighting factor α to 0.5 implies the codebooks and the whole-utterance frames are equally treated without bias. Although $\alpha = 0.5$ is not necessarily an optimal assignment, at least it implies that there is little need for meticulous tuning of the weighting factor α in order to obtain an improved performance for these associative methods.

5. A particular phenomenon for these associative methods (A-CMS, A-CMVN, and A-HEQ) is that increasing the number of clean-speech codewords R does not improve the recognition accuracy, which somewhat contradicts the results for codebook-based CMS and CMVN, as shown in Tables 1 and 2. For example, increasing the value of R from 16 to 256 in A-HEQ results in an accuracy degradation of 1.40%. One of the possible reasons for this degradation is the inconsistency between the codebooks and the whole-utterance frames in these associative methods. The codebooks present only the characteristics of the speech portions in the utterance, while the whole utterance contains both speech and non-speech portions. Increasing the codebook size somewhat portrays the speech characteristics more precisely, thus, highlights the above inconsistency further. From the viewpoint of implementation, however, it becomes an advantage since we can obtain better recognition results with fewer codewords in these associative methods, which reduces the computation complexity.

Table 3. Recognition accuracy (%) achieved by utterance-based, codebook-based, and associative CMS averaged across the SNRs between 0 and 20dB, under the complex back-end structure. AR (%) and RR (%) are the absolute and relative error rate reductions over the baseline.

Method	Set A	Set B	Average	AR	RR
Baseline	71.92	67.79	69.86	—	—
U-CMS	79.37	82.47	80.92	11.07	36.71
C-CMS ($R = 256$)	81.46	81.49	81.48	11.62	38.55
A-CMS ($R = 16, \alpha = 0.5$)	83.28	84.92	84.10	14.25	47.25
A-CMS ($R = 64, \alpha = 0.5$)	82.92	84.89	83.91	14.05	46.61
A-CMS ($R = 256, \alpha = 0.5$)	82.13	84.29	83.21	13.36	44.30

Table 4. Recognition accuracy (%) achieved by utterance-based, codebook-based, and associative CMVN averaged across the SNRs between 0 and 20dB, under the complex back-end structure. AR (%) and RR (%) are the absolute and relative error rate reductions over the baseline.

Method	Set A	Set B	Average	AR	RR
Baseline	71.92	67.79	69.86	—	—
U-CMVN	85.03	85.56	85.30	15.44	51.22
C-CMVN ($R = 256$)	87.10	87.32	87.21	17.36	57.57
A-CMVN ($R = 16, \alpha = 0.5$)	87.87	88.67	88.27	18.42	61.09
A-CMVN ($R = 64, \alpha = 0.5$)	87.34	88.24	87.79	17.94	59.50
A-CMVN ($R = 256, \alpha = 0.5$)	87.25	88.06	87.66	17.80	59.05

Table 5. Recognition accuracy (%) achieved by utterance-based, codebook-based, and associative HEQ averaged across the SNRs between 0 and 20dB, under the complex back-end structure. AR (%) and RR (%) are the absolute and relative error rate reductions over the baseline.

Method	Set A	Set B	Average	AR	RR
Baseline	71.92	67.79	69.86	—	—
U-HEQ	86.95	88.39	87.67	17.82	59.10
C-HEQ ($R = 256$)	86.23	85.77	86.00	16.15	53.56
A-HEQ ($R = 16, \alpha = 0.5$)	90.21	91.16	90.69	20.83	69.10
A-HEQ ($R = 64, \alpha = 0.5$)	88.93	89.68	89.31	19.45	64.52
A-HEQ ($R = 256, \alpha = 0.5$)	88.84	89.73	89.29	19.43	64.46

Figures 2, 3, and 4 show the averaged recognition accuracy rates for each of the eight noise conditions in Test Sets A and B achieved by various types of CMS, CMVN, and HEQ. Roughly speaking, the four noise types, "subway," "street," "car," and "exhibition" can be viewed as stationary noise, while the other four noise types, "restaurant," "babble," "airport," and "train-station" are non-stationary noise. From the three figures, it is first found that, the utterance-based methods perform better in the non-stationary noise cases than in the stationary noise cases, while the situation is reversed for codebook-based methods. Second, the accuracy variation due to different noise conditions is more significant in the utterance-based and codebook-based methods than in the associative methods. Third, for each noise type, the associative method always performs better than the corresponding utterance-based and codebook-based ones, which again indicates that integrating the utterance and codebook information in these feature statistics methods is quite helpful for a wide range of noise

environments.

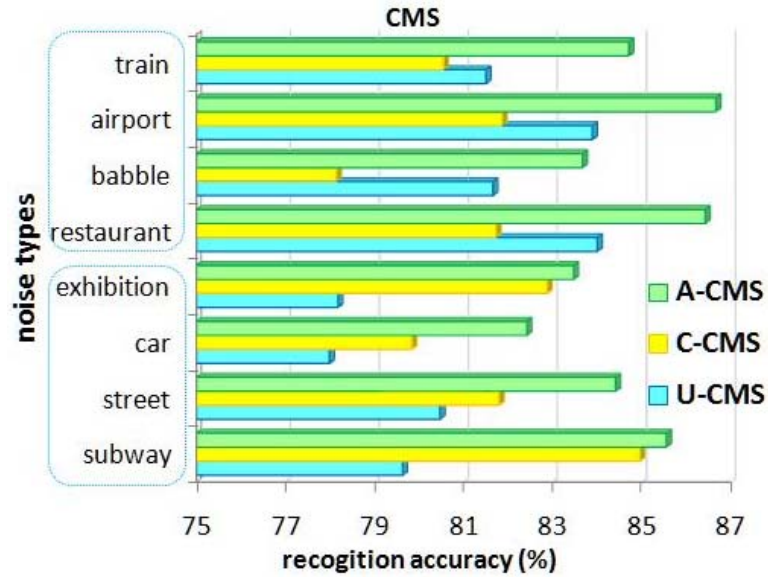

Figure 2. Recognition accuracy (%) achieved by three CMS methods, U-CMS, C-CMS (R=256), and A-CMS (R=16, α=0.5) for eight noise types in Test Sets A and B, averaged over five SNR conditions, 0~20dB.

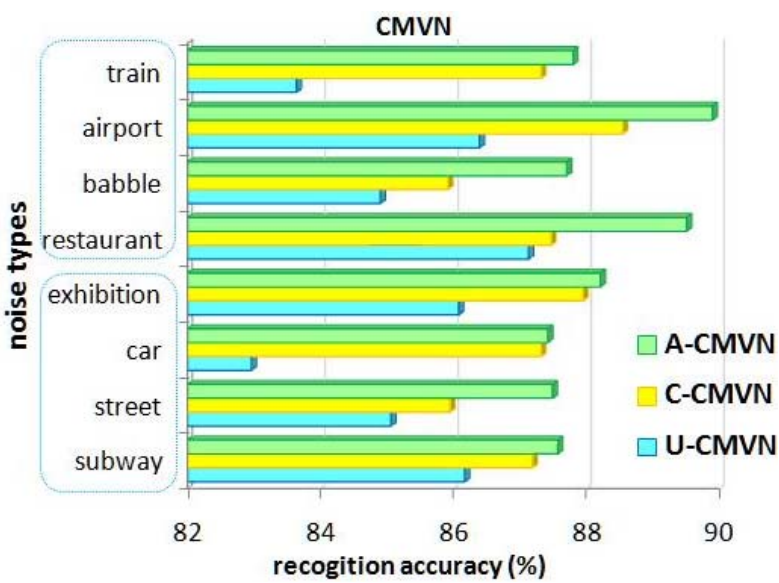

Figure 3. Recognition accuracy (%) achieved by three CMVN methods, U-CMVN, C-CMVN (R=256), and A-CMVN (R=16, α=0.5) for eight noise types in Test Sets A and B, averaged over five SNR conditions, 0~20dB.

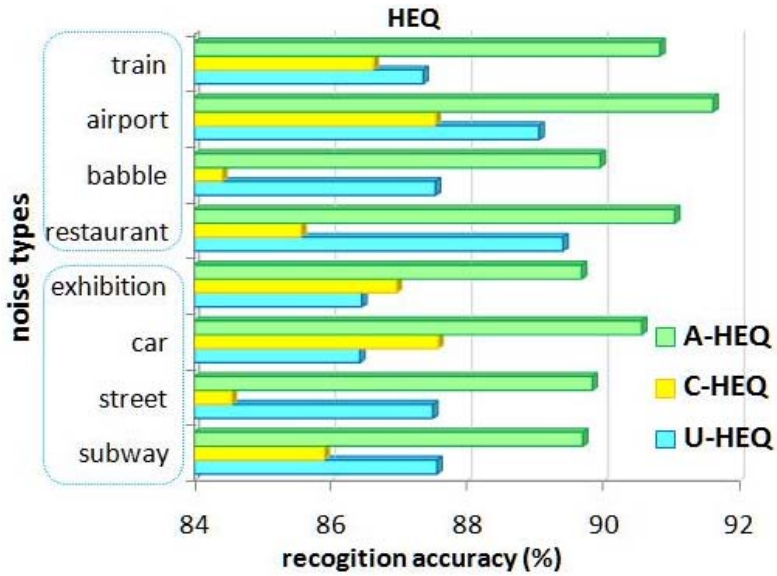

Figure 4. Recognition accuracy (%) achieved by three HEQ methods, U-HEQ, C-HEQ (R=256), and A-HEQ (R=16, α=0.5) for eight noise types in Test Sets A and B, averaged over five SNR conditions, 0~20dB.

5.3 The Effect of the Weighting Factor α in the Associative Methods

Here, the effect of the weighting factor α on the proposed associative methods is investigated. As stated in Section 2, the weighting factor α determines the usage ratio between the codebooks and the whole-utterance frames in estimating the feature statistics. Here, the size R of the clean speech codebook is fixed at 16 since it brings the best recognition accuracy, as mentioned in the previous subsection. Then, different assignments of the weighting factor α from 0 to 1, spaced at 0.1 intervals, are given for A-CMS, A-CMVN, and A-HEQ.

Figures 5, 6, and 7 show the recognition results averaged over five SNR conditions (0~20dB) and all eight noise types in Test Sets A and B for different values of α for A-CMS, A-CMVN, and A-HEQ, respectively. Note that these associative methods with $\alpha = 0$ and $\alpha = 1$ behave equally to the corresponding utterance-based and the codebook-based methods, respectively. For example, A-HEQ with $\alpha = 0$ is identical to U-HEQ, and A-HEQ with $\alpha = 1$ is identical to C-HEQ. From the three figures, we first find that, with any value of α that is not equal to 0 or 1, the newly-proposed associative methods always behave better than both the utterance-based and codebook-based ones. This result again supports our previous comment that integrating both codebook and utterance knowledge promotes the performance of feature statistics normalization techniques. Next, for different associative

methods, the corresponding optimal α values that achieve the optimal performance are not identical to each other. For example, the optimal α for A-CMS, A-CMVN, and A-HEQ are 0.3, 0.8, and 0.4, respectively. For each method, however, the differences among the accuracy rates obtained with different α are in fact relatively slight when α is in the range $[0.3, 0.8]$ (*i.e.*, $0.3 \le \alpha \le 0.8$). The maximum deviation in the accuracy rates for A-CMS with varying α is 1.69%, and it is 1.08% and 0.70% for A-CMVN and A-HEQ, respectively. Furthermore, setting $\alpha = 0.5$ just results in the accuracy degradation of 0.85%, 0.49%, and 0.01% for A-CMS, A-CMVN, and A-HEQ, respectively, when compared with the optimal values. This result implies that we just have to evenly employ the codebooks and the whole-utterance frames in these associative methods, then the nearly optimal performance can be achieved.

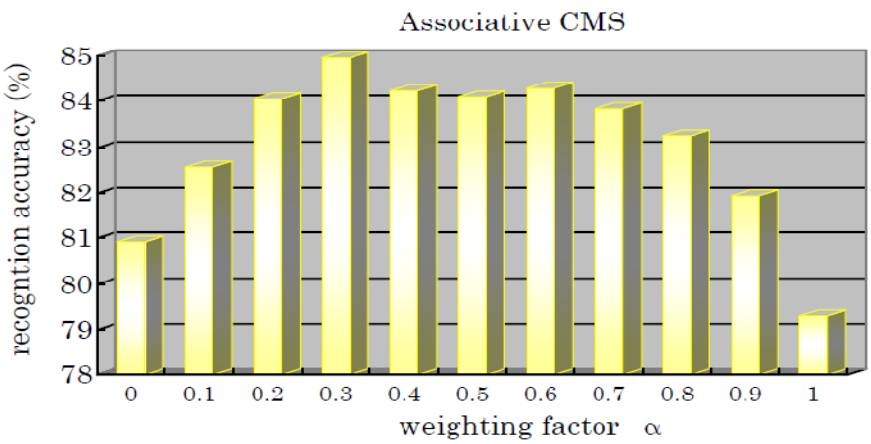

Figure 5. Recognition accuracy (%) averaged over five SNR values and all the eight noise types in Test Sets A and B vs. different assignments of the weighting factor α in associative CMS.

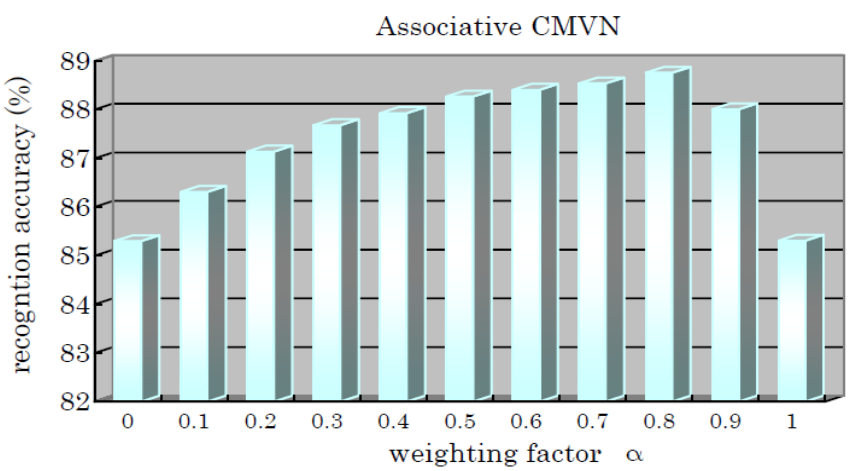

Figure 6. Recognition accuracy (%) averaged over five SNR values and all the eight noise types in Test Sets A and B vs. different assignments of the weighting factor α in associative CMVN.

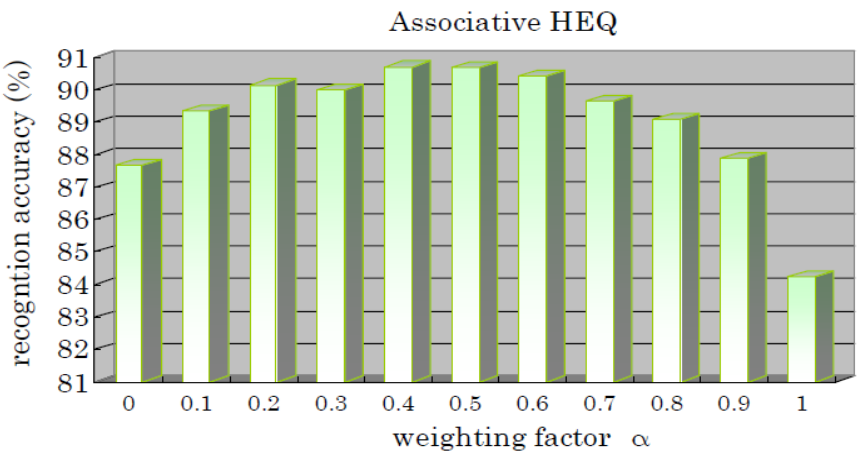

Figure 7. Recognition accuracy (%) averaged over five SNR values and all the eight noise types in Test Sets A and B vs. different assignments of the weighting factor α in associative HEQ.

5.4 Comparison of the Associative Methods with the Other Noise-Robust Techniques

In the previous subsections, we have shown that the associative methods outperform the corresponding utterance-based and codebook-based methods. Here, the associative methods are compared with the other two noise-robust techniques: mean-and-variance normalization plus autoregressive moving average (ARMA) filtering (MVA) (Chen & Bilmes, 2007) and the ETSI advanced front-end (AFE) feature extraction algorithm (ETSI standard doc, 2003). In MVA, an ARMA filter is performed on the (utterance-based) MVN-processed features in order to emphasize the relatively low modulation frequency components. On the other hand, the AFE makes use of a two-stage Wiener filter in order to reduce noise. For the purpose of comparison, we also perform the low-pass ARMA filtering on the A-CMVN processed features, which is called A-MVA here.

Figures 8 and 9 present the recognition accuracy rates of the various approaches for Test Sets A and B, respectively. From the two figures, we have the following observations:

1. Both MVA and AFE perform very well. The superior performance of MVA over U-CMVN shows that the low-pass ARMA filter helps extract the noise-robust components in U-CMVN processed features. On the other hand, AFE performs the best among all the methods here, which implies that, in AFE, the two-stage Wiener filtering process achieves very effective noise reduction.

2. A-CMVN behaves as well as MVA (U-CMVN plus ARMA), and in A-MVA the low-pass ARMA filtering process offers A-CMVN an accuracy improvement of 1.76% and

1.64% for Test Sets A and B, respectively. This result shows that, similar to U-CMVN, A-CMVN is additive to the ARMA filtering to provide better performance.

3. A-HEQ performs worse than AFE by 2.13% and 0.83% for Set A and Set B, respectively, in recognition accuracy. The possible explanation is: the noise estimates are quite accurate in AFE, which benefit the two-stage Wiener filtering process a lot, while A-HEQ just employs the several leading frames in an utterance as the noise estimates. As a result, in our future work, we will attempt to incorporate the noise estimates in AFE with the proposed associative methods in order to enhance their noise-robustness capability.

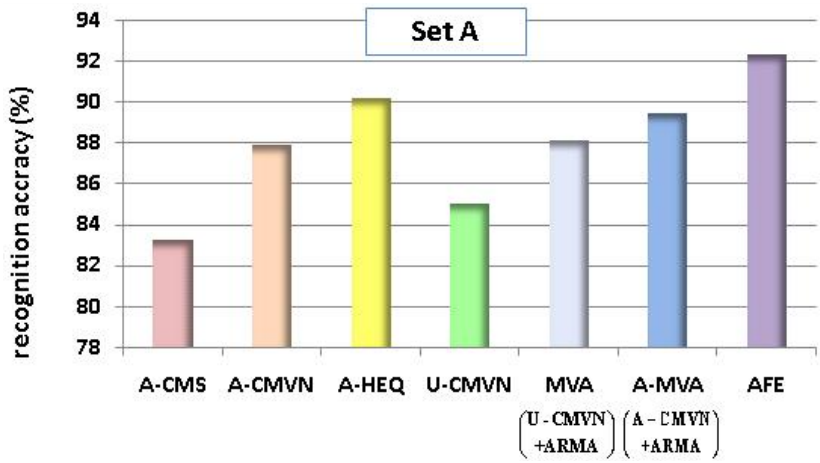

Figure 8. Recognition accuracy (%) achieved by various approaches averaged over five SNR values and all the four noise types in Test Set A, under the complex back-end structure.

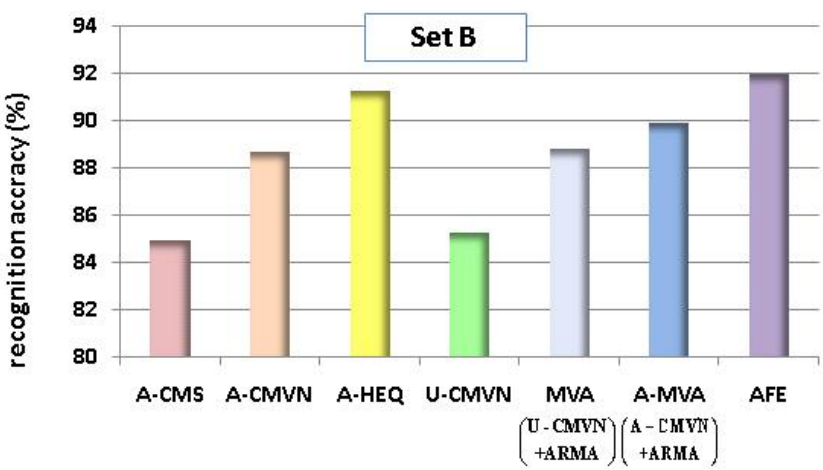

Figure 9. Recognition accuracy (%) achieved by various approaches averaged over five SNR values and all the four noise types in Test Set B, under the complex back-end structure.

5.5 Experimental Results of Our Proposed Methods with the Standard Back-end Defined in Aurora-2 Database

The recognition experiments in the previous four subsections use a more complicated hidden Markov model (HMM) structure, which follows the Microsoft advanced back-end training scripts (Droppo, Deng, & Acero, 2002), and each digit HMM has 16 states and 20 mixtures per states. Here, we train each digit HMM as 16 states and 3 mixtures per states following the standard back-end training scripts (Hirsch & Pearce, 2000), and the corresponding recognition results for our proposed methods are shown in Table 6. Comparing Table 6 with Tables 3, 4, and 5, we have the following observations:

1. Under the simpler HMM structure, the recognition accuracy rates achieved by each method become worse, which implies that a small number of mixtures cannot adequately represent the short-term speech characteristics.

2. The accuracy difference between the utterance-based and codebook-based methods becomes more significant. For example, C-CMS and C-CMVN outperform U-CMS and U-CMVN by 6.62% and, 6.71%, respectively (in Tables 3 and 4, the accuracy differences are less than 2%), and U-HEQ outperforms C-HEQ by 6.36% (in Table 5, the accuracy difference is just 1.67%).

3. In almost all cases, the proposed associative methods perform better than the corresponding utterance-based and codebook-based methods. Therefore, it reveals that, regardless of the recognition model complexity, integrating the codebook and utterance information helps enhancing the feature statistics normalization methods and thus brings about better recognition performance under additive noise environments.

Table 6. Recognition accuracy (%) achieved by various approaches averaged across the SNRs between 0 and 20dB, under the standard back-end structure. AR (%) and RR (%) are the absolute and relative error rate reductions over the baseline.

Method	Set A	Set B	Average	AR	RR
Baseline	61.97	55.78	58.88	—	—
U-CMS	64.36	69.43	66.90	8.02	19.50
C-CMS ($R = 256$)	72.40	74.64	73.52	14.64	35.60
A-CMS ($R = 16, \alpha = 0.5$)	73.72	77.51	75.61	16.73	40.69
U-CMVN	73.83	75.01	74.42	15.54	37.79
C-CMVN ($R = 256$)	80.88	81.39	81.13	22.25	54.11
A-CMVN ($R = 16, \alpha = 0.5$)	80.74	81.75	81.24	22.36	54.38
U-HEQ	80.33	81.24	80.79	21.91	53.28
C-HEQ ($R = 256$)	75.21	73.64	74.43	15.55	37.82
A-HEQ ($R = 16, \alpha = 0.5$)	82.96	83.85	83.41	24.53	59.65

6. Concluding Remarks and Future Works

In this paper, we propose associative CMS, CMVN, and HEQ, in which the required statistical information is obtained by incorporating advanced pseudo stereo codebooks and the processed whole-utterance frames. These new approaches demonstrate enhanced robustness of speech features under various additive noise environments. Compared with conventional utterance-based and codebook-based CMS, CMVN, and HEQ, these new approaches provide significantly better recognition performance. In our future work, we will employ the proposed statistics evaluation scheme to other feature statistics normalization approaches, like CGN (Yoshizawa, Hayasaka, Wada, & Miyanaga, 2004), HOCMN (Hsu & Lee, 2004), and CSN (Du & Wang, 2008), to investigate if better performance can be achieved.

References

Acero, A. (1990). *Acoustical and environmental robustness in automatic speech recognition*. Ph.D. dissertation, Department of Electrical and Computer Engineering, Carnegie Mellon University, Pittsburg, PA.

Acero, A., Deng, L., Kristjansson, T., & Zhang, J. (2000). HMM adaptation using vector Taylor series for noisy speech recognition. In *Proceeding of 2000 International Conference on Spoken Language Processing (ICSLP 2000)*, 3, 869-872.

Atal, B.S. (1974). Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification. *Journal of the Acoustic Society of America*, 55, 1304-1312.

Beattie, V. L., & Young, S. J. (1992). Hidden Markov model state-based cepstral noise compensation. In *Proceeding of International Conference on Spoken Language Processing (ICSLP 1992)*, 519-522.

Berstein, A. D., & Shallom, I. D. (1991). An hypothesized Wiener filtering approach to noisy speech recognition. In *Proceeding of 1991 International Conference on Acoustics, Speech and Signal Processing (ICASSP 1991)*, 2, 913-916.

Boll, S.F. (1979). Suppression of acoustic noise in speech using spectral subtraction. *IEEE Trans. on Acoustics, Speech and Signal Processing*, 27(2), 113-120.

Chen, C.-P., & Bilmes, J. A. (2007). MVA Processing of Speech Features. *IEEE Trans on Audio, Speech, and Language Processing*, 15(1), 257-270.

Deng, L., Acero, A., Jiang, L., Droppo, J., & Huang, X. (2001). High-performance robust speech recognition using stereo training data. In *Proceeding of 2001 International Conference on Acoustics, Speech and Signal Processing (ICASSP 2001)*, 1, 301-304.

Droppo, J., Deng, L., & Acero, A. (2001). Evaluation of the SPLICE Algorithm on the Aurora2 Database. In *Proceeding of 2001 Eurospeech Conference on Speech Communications and Technology (Eurospeech 2001)*, 217-220.

Droppo, J., Deng, L., & Acero, A. (2002). Evaluation of SPLICE on the AURORA 2 and 3 tasks. In *Proceeding of 2002 International Conference on Spoken Language Processing (Interspeech 2002－ICSLP)*, 29-32.

Du, J. & Wang, R.-H . (2008). Cepstral shape normalization (CSN) for robust speech recognition. In *Proceeding of 2008 International Conference on Acoustics, Speech and Signal Processing (ICASSP 2008)*, 4389-4392.

ETSI standard doc. (2003). Speech Processing, transmission and quality aspects (STQ); Distributed speech recognition; Advanced feature extraction algorithm. ETSI ES 202 050 v1.1.3 (2003-11).

Gales, M. J. F. & Young, S. J. (1993). Cepstral parameter compensation for HMM recognition in noise. *Speech Communication*, 12, 231-239.

Gales, M. J. F. & Young, S. J. (1995a). Robust speech recognition in additive and convolutional noise using parallel model combination. *Computer Speech and Language*, 9, 289-307.

Gales, M. J. F. & Young, S. J. (1995b). A fast and flexible implementation of parallel model combination. In *Proceeding of 1995 International Conference on Acoustics, Speech and Signal Processing (ICASSP 1995)*, 133-136.

Hermansky, H. (1990). Perceptual linear predictive (PLP) analysis of speech. *Journal of the Acoustical Society of America*, 87(4), 1738-1752.

Hilger, F. & Ney, H. (2006). Quantile based histogram equalization for noise robust large vocabulary speech recognition. *IEEE Trans. on Audio, Speech and Language Processing*, 14(3), 845-854.

Hirsch, H. G. & Pearce, D. (2000). The AURORA experimental framework for the performance evaluations of speech recognition systems under noisy conditions. In *Proceeding of ISCA ITRW ASR2000 "Automatic Speech Recognition: Challenges for the Next Millennium*", Paris, France, 181-188.

Holmes, J. N. & Sedgwick, N. C. (1986). Noise compensation for speech recognition using probabilistic models. In *Proceeding of 1986 International Conference on Acoustics, Speech and Signal Processing (ICASSP 1986)*, 11, 741-744.

Hsu, C.-W. & Lee, L.-S. (2004). Higher order cepstral moment normalization (HOCMN) for robust speech recognition. In *Proceeding of 2004 International Conference on Acoustics, Speech and Signal Processing*, 197-200.

Hung, J.-W. (2006). Cepstral statistics compensation using online pseudo stereo codebooks for robust speech recognition in additive noise environments. In *Proceeding of 2006 International Conference on Acoustics, Speech and Signal Processing (ICASSP 2006)*.

Hung, J.-W. (2008). Cepstral statistics compensation and normalization using online pseudo stereo codebooks for robust speech recognition in additive noise environments. *IEICE Trans. on Information and Systems*, E91-D(2), 296-311.

ITU recommendation G.712. (1996). Transmission performance characteristics of pulse code modulation channels. Nov. 1996.

Klatt, D. H. (1979). A digital filterbank for spectral matching. In *Proceeding of 1979 International Conference on Acoustics, Speech and Signal Processing (ICASSP 1979)*, 573-576.

Lee, C.-H. (1998). On stochastic feature and model compensation approaches to robust speech recognition. *Speech Communication*, 25, 29-47.

Leggester, C. J. & Woodland, P. C. (1995). Maximum likelihood linear regression for speaker adaptation of continuous density HMMs. *Computer Speech and Language*, 9, 171-186.

Makhoul, J. (1975). Spectral linear prediction: properties and applications. *IEEE Trans. on Acoustics, Speech and Signal Processing*, 23(3), 283-296.

Moreno, P. J., Raj, B., & Stern, R. M. (1996). A vector Taylor series approach for environment-independent speech recognition. In *Proceeding of 1996 International Conference on Acoustics, Speech and Signal Processing (ICASSP 1996)*, 733-736.

Moreno, P. J., Raj, B., & Stern, R. M. (1998). Data-driven environmental compensation for speech recognition: A unified approach. *Speech Communication*, 24(4), 267-285.

Nadas, A., Nahamoo, D., & Picheny, M. (1988). Speech recognition using noise-adaptive prototypes. In *Proceeding of 1988 International Conference on Acoustics, Speech and Signal Processing (ICASSP 1988)*, 517-520.

Sankar, A. & Lee, C.-H. (1996). A maximum-likelihood approach to stochastic matching for robust speech recognition. *IEEE Trans. on Acoustics, Speech and Signal Processing*, 4(3), 190-202.

Segura, J. C., Benitez, M. C., de la Torre, A., Dupont, S., & Rubio, A. J. (2002). VTS residual noise compensation. In *Proceeding of 2002 International Conference on Acoustics, Speech and Signal Processing (ICASSP 2002)*, 1, I-409-I-412.

Tai, C-F. & Hung, J-W. (2006). Silence energy normalization for robust speech recognition in additive noise environments. In *Proceeding of 2006 International Conference on Spoken Language Processing (Interspeech 2006 ─ICSLP)*, 2558-2561.

Tibrewala, S. & Hermansky, H. (1997). Multiband and adaptation approaches to robust speech recognition. In *proceeding of 1997 Eurospeech Conference on Speech Communications and Technology (Eurospeech 1997)*, 2619-2622.

Varga, A. P. & Moore, R. K. (1990). Hidden Markov model decomposition of speech and noise. In *Proceeding of 1990 International Conference on Acoustics, Speech and Signal Processing (ICASSP 1990)*, 845-848.

Yoshizawa, S., Hayasaka, N., Wada, N., & Miyanaga, Y. (2004). Cepstral gain normalization for noise robust speech recognition. In *Proceeding of 2004 International Conference on Acoustics, Speech and Signal Processing (ICASSP 2004)*, I-209-212.