# Context-Centered Template Matching for Chinese Lexicon Construction

*GUO JIN*
Institute of Systems Science
National University of Singapore
Kent Ridge, Singapore 119597
email: guojin@iss.nus.sg

## *ABSTRACT*

Our objective is to develop productive approaches for constructing million-entry Chinese lexicon. The emphasis in this paper is on the development of the fresh context-centered lexicon construction methodology. Beginning with an intuitive explanation, we gradually worked out the formal mathematical model and the productive lexicon construction algorithm. The effectiveness of the proposed scheme is demonstrated in a detailed case study.

*KEYWORDS:* Lexicon Construction, Term Identification, Chinese Computing, Language Modeling, Statistical and Corpus-based NLP.

## 1 INTRODUCTION

The objective in this paper is to build a lexicon with huge number of entries. The task is strongly application motivated. For example, the Chinese Dictation Kit (CDK) developed at Apple-ISS (Yuan, etc., 1996) is based on a lexicon of over 350,000 entries. The Chinese-English Machine Translation System from SYSTRAN (Yang and Gerber, 1996) employs a lexicon of over 600,000 words. From my personal experiences in Chinese text spell checking and proofreading (Guo, 1994) and Pinyin-to-Hanzi transcription (Guo, 1993), huge lexicon on the size of million entries would be of great help, especially for alternative suggestion and error correction. Large lexicon can also be found in many other applications such as text-to-speech (Sproat, 1994) and information retrieval (Harman, 1995, Harman, 1996).

Moreover, such lexicon is largely for general purpose rather than domain specific. For example, I was told that most obvious domain specific and/or technical terms are purposely excluded from the Apple's CDK lexicon. This is to ensure the general suitability of the system. Domain specific entries are supported with a function limited supplementary user dictionary management mechanism.

Inevitably, many entries in such huge lexicon are not simple words but compounds, idioms, rigid phrases, or even complete (short) sentences. The question here is not the justification of whether or not entries in such lexicon are words, or such huge collection is still a lexicon proper. Rather, what is challenging us is: how do we build up such a huge lexicon and make it useful?

Past efforts in unknown word identification and dictionary construction can be classified into the following categories. Manual construction is of course the first way. The 600,000 words SYSTRAN lexicon (Yang and Gerber, 1996) is from a government department. Automatic extraction from large (partially) parsed corpora such as Brown corpus (Kucera and Francis, 1967) and the Penn TreeBank (Markus, et al., 1993) also falls into this category. This is definitely a quality approach, provided that enough resources are available.

Grammar-based automatic generation is the second way. Basically, the 350,000 words Apple CDK lexicon (Yuan, etc., 1996) is from the enumeration of some well-designed ***Prefix-Body-Suffix*** patterns on a core dictionary of about 10,000 multi-character words. For example, as an entry, the Chinese phrase ***"很美丽的 (very beautiful)"*** follows the pattern ***"adv + adj + de"*** where "很 (very)" is the adverbial prefix, "美丽 (beauty)" the body and "的 (de mark)" the suffix sometimes functioned

similar to the English adjective suffix "-ful". One of the obvious problems is the lexicon's (lack of) coverage. Too many words and terms are not compositional, at least not composed in such a simple way.

Sub-language modeling is the third way. For example, Sun and Huang (1996) systematically presented several intelligent ***agents*** each for identifying a type of constructs such as Chinese names (CName Agent), Transcribed Foreign Names (TFName Agent) and Chinese Place Names (CPName Agent). Song, etc., (1996) presented sets of rules for company name and person's name identification. Mo, etc., (1991) derived the grammar for Chinese-specific determination-measure compounds and implemented an identification parser. Chen and Liu (1992) implemented rules for reduplication and A-not-A constructions and some other kinds of derivable constructions. Sub-language approach is the mainstream in unknown word processing in Chinese text processing (Chang, 1994, Nie, Jin and Hannan, 1994, Zheng and Liu, 1993, Lee, Lee and Chen, 1994, among many others). All these works are limited to their predefined type of constructs.

Entity-centered statistical modeling is yet another way. The assumption behind this school is that "words are tightly-bounding and frequently-using entities". Fung and Wu (1994) applied ***CXtract***, a localized version of ***Xtract*** by Smadja (1993), to extract statistically significant Chinese character ngrams as possible entries for dictionary augmentation. This approach is not applicable for the identification of the majority of low frequency words, a drawback deeply rooted in its underlying assumption (Smadja, 1993)*. From a corpus of about 2 million Chinese characters, Fung and Wu (1994) only extracted about 5,000 new entities. Other examples of works in this category are Chiang, etc., (1992), Lin, etc., (1993) and Sproat and Shih (1990). ***Mutual information*** and ***t-test*** are the two representative means in this school (Church and Hanks, 1990).

The only work I am aware of which is more or less away from the above mentioned categories is by Luk (1995). He proposed to use his heuristically defined "lexicographic index" as an indicator to highlight potential "words". The unique characteristics of his approach is its capability of extracting low frequency words: a character string is highlighted as long as it occurred in at least two different contexts. With this approach, Luk (1996) extracted 75,535 new entities from the 4-million-character PH corpus (Guo and Lui, 1992), clearly higher recall than what Fung and Wu (1994) achieved with their entity-centered ngram statistical approach, if the two are comparable.

To reach the million-entry target, however, the recall rate must be significantly higher. As Luk (1995) has already loosed the requirement to two different occurs, the question here basically becomes: ***how do we extract words that occur only ONCE in corpus***?

Note, even for huge corpus, the majority are still those occurring few times. For instance, in a collection of about 60 million characters news articles from China's Xinhua News Agency, there are 1,129,313 unique word bigram types, of which 339,839 bigram types occur once, and 172,467 twice. That is, even for a corpus of that huge size, there are still about half bigram types occurring merely once or twice. That is yet a result based on close-dictionary tokenization (that is, there is no effort taken for unknown word identification in the process of text tokenization and all tokenized words are in the given tokenization dictionary). Situations for trigrams and general ngrams under open-dictionary tokenization will be even more apparent.

Some readers might argue that those low frequency entities are not important, simply because each of them, as an individual entity, is next to ***never been used***. That is true in one sense. Collectively, however, the "silent majority" forms a large mass. If all bigram types occurring less than 11 times are filtered out as Fung and Wu (1994) did, on average, there will be at least one unknown bigram type in each sentence†. As bigrams are the core of unknown words or compounds, discarding low frequency

---

* Smadja (1993, page 165) himself pointed out that: "***Xtract*** has only been effective at retrieving collocations for words appearing at least several dozen times in the corpus". However, "For the 10 million-word stock market corpus, there are some 60,000 different word forms" and "Out of the 60,000 words in the corpus, only 8,000 were repeated more than 50 times."

† This could also be validated on data from many other publications. For example, Church and Gale (1991) listed in their Table 1 that, in a corpus of 22 million words, bigram types with frequency up to 9 totally count for 5,210,157 occurrences. That is, on average, for every four bigrams encountered, there is a low-frequency bigram type. And normally sentence length is far more than four words.

bigram types effectively ruled out the possibility of collecting and utilizing that significant portion of knowledge.

Entity-centered statistical approaches are good for highlighting those few "top stars". In contrast, our attention is on the largely ignored "silent majority". The key idea here is to let context play the center role.

We will first in Section 2 give an intuitive explanation of our basic idea, and then, in Section 3, present a detailed case study. Section 4 is reserved for formal presentation of our mathematical modeling. Then, the baseline lexicon construction algorithm is given in Section 5. Then a short conclusion is in Section 6. Complete data list for the case study is in Appendix.

## 2 INTUITION

In this section, we will first give an intuitive start of our context-centered thinking, then contrast it with traditional entity-centered thinking, and sketch our context-centered lexicon augmentation algorithm.

### 2.1 "This is a word."

Suppose I have no idea of English but was told that the sentences below are all legitimate in English:

(1.a) This is a ***pear***.
(1.b) This is a ***table***.
(1.c) This is a ***man***.
(1.d) This is a ***dog***.

Moreover, I was told that ***pear***, ***table***, ***man*** and ***dog*** are all legitimate English ***words***. Then, if I know the sentence

(1.e) This is a ***xxx***.

is also correct, by analogy, I will feel comfortable to agree that the entity ***xxx*** here ought to be an English ***word*** also. Moreover, I am happy to accept the ***assertion*** that: ***any single entity taking the position of "xxx" in the sentence above is a*** **word**. In other words, any entity which could be filled up into the empty slot in

(2) This is a _____.

is a word. In this paper, the empty slot is named a ***word holding slot*** or simply ***(word) slot*** and (part of) the sentence containing such word holding slot a ***word holding template*** or simply ***(word) template***.

To go a step further, let us agree that "***a pear***", "***a table***", "***a man***" and "***a dog***" are all a type of entity called ***chunk*** (Abney, 1991, Abney, 1996). Notice that all these chunks take the empty slot in sentence:

(3) This is _____.

Then, it should sound reasonable to accept that: ***anything taking the empty slot above is a chunk***. Or, the empty slot is a ***chunk holding slot*** and (part of) the sentence a ***chunk holding template***.

### 2.2 Context-centered Definition

Note, we do define what are words or chunks above. Explicitly, ***a word is an entity that could take up a word holding slot in a word holding template.***

This definition makes itself apart from the tradition of entity-centered thinking. Linguistically, as evidenced in grammar books and dictionaries, "word" is normally defined as an entity with "phonological, lexicographic, syntactic and/or semantic independence (meaning)", and "free usage in text" (e.g., Hu, 1987). In NLP community, the widely followed criterion for justifying word is the notorious eight Chinese characters: ***"结合紧密, 使用稳定 (tight-in-combination and stable-in-use)"*** (GB-13715, 1993). We call them ***entity-centered definitions*** since the judgment is mostly based on the characteristics of the entity to be judged, but with no explicit reference to context.

Such entity-centered definitions worked fine for those high frequency "top stars". If an entity occurred thousand or even million times in text, it is easy to judge whether or not it is "tight-in-combination

and stable-in-use" or "with phonological, lexicographic, syntactic and/or semantic independence (meaning)". However, such high frequency entities, if they *are* words, are by and large already recorded in out-of-the-shelf dictionaries, a resource already in use. What becomes problematic are exactly those with low usage frequencies. Yet, without high profile in text, we are not able to reliably determine their (statistical) characteristics.

To the "silent majority", context becomes the primary source. Yes, as long as you believe in the correctness of the sentence "This is a xxx", you will not hesitate to agree with the acceptability of "xxx" as a word, no matter how rare, odd, exotic, strange or mysterious it is. We only need to see *one* occurrence of the word and confidently make our judgment. That is the power of context, an information source largely neglected in literature.

Note, we are not rejecting the entity-centered thinking. What we are emphasizing here is that, with the compensation of the context-centered thinking, we may even have a more complete understanding. The entity-centered model and the context-centered model are somehow in opposition. Both have their pros and cons. Yet they could co-exist in one system, as they are the complement of each other in nature. In this paper I will purposely ignore the good part of the entity-centered model.

### 2.3 Algorithm Sketch

The intuitive illustration above suggests itself the following two-phase dictionary augmentation *algorithm*:

***Phase 1: Template Preparation***

This is to collect a large set of high quality word holding templates such as "This is a ___.". Later, we will discuss how this could be done in an automatic manner on corpus and dictionary.

***Phase 2: Template Matching***

This is to identify words from text by matching sentences against above prepared templates. Those entities in text with surrounding context matching some templates and themselves taking word holding slots will be collected as candidates for dictionary augmentation.

The core of our dictionary augmentation algorithm is that simple. To actually put the algorithm into practical use, of course, some preprocessing such as dictionary-based tokenization and part-of-speech tagging, and some postprocessing such as linguistic filtering and statistical selection, are also required. Our focus, however, is only on the above-mentioned two phases. Before going into details of the two phases, let us have a real world case study first.

## 3 CASE STUDY

The task in this section is to have a detailed case study on context-centered template matching word identification. We will first introduce our template pattern, and discuss various properties of a particular template used in this case study. Then, word candidates extracted from corpus with that particular template are presented, categorized and analyzed. We will also make a short conclusion in the last subsection.

### 3.1 Template Pattern

Practically, to make the collection of word holding templates manageable, most word templates could not be full sentences but small sentence fragments. Although various template types are possible, we only tried the simplest one as given below:

```
<LeftContext> <WordHoldingSlot> <RightContext>
```

Depending on the level of preprocessing, the `<LeftContext>` and `<RightContext>` could be strings of characters (for un-tokenized text), words (for tokenized text), part-of-speeches (for part-of-speech tagged text), or their combinations. Similarly, the `<WordHoldingSlot>` could take up a fixed or variable number of characters, words and/or part-of-speeches.

## 3.2 <发展> <character bigram> <comma>

To be specific, let us examine in detail a relatively simple real world example given below.

<发展> <character bigram> <comma>

This template is exactly five characters long, with the left two characters fixed to the two Chinese characters "发展 (develop or development, if used as a word)", and the rightmost character bounded to the specific punctuation mark comma. The middle two character positions are left unconstrained to form the word holding slot. For this template, there is no text preprocessing required.

This template is not as trivial as the template "This is a ___." we illustrated in the section above. It is purposely selected for highlighting some potential inherent difficulties.

First, we assume no text pre-tokenization. At least theoretically, this will bring in all kinds of traps associated with the notorious Chinese text tokenization problem widely believed to be caused by the lack of English blank space equivalent word delimiter in Chinese text. In Chinese, "发展 (develop, development)" itself is a legitimate word, but "发展中 (developing)" is also an entry in some dictionaries. Moreover, in sentence "张发展翅飞翔 (Zhang Fa is flying high)", the character "发 (Fa)", the first character in "发展", is in fact the given name of the person "张发 (Zhang Fa)" and thus a part of the proper noun, and "展", the second character in "发展", is the initial character of the predicative idiom "展翅飞翔 (fly high)". In short, depending on the context it appeared, the continuous character bigram "发展" could be used as a word, or part of a word, or parts of two adjacent words.

Even if the text is properly tokenized and "发展" is confirmed to be used as a word in text, its part-of-speech is nevertheless ambiguous. Chinese is a language lack of inflections. Depending on the context it is used, the two-character word "发展" could be used as either a noun ("development") or a verb ("develop"), yet written exactly the same. We have to count on the context for part-of-speech determination and/or disambiguation. But in the template, in terms of the part-of-speech ambiguous Chinese word "发展", its right context is by design a "word holding slot" meaning an unknown word, yet its left context is not mentioned at all.

In addition, the template is not a full sentence. We do not even know whether or not the five-character-long template is a syntactically or semantically self-contained unit like a phrase or a term. Rather, in terms of the word holding slot, only two left characters and one right character are given. That is the only indicator or constraint.

In a word, the information provided in this template is rather poor. Many ambiguities and unknowns could be expected to arise even for human expert explorers simply because of the lack of information. Realistically, we could not expect high predictability of the template in word recognition.

Nevertheless, not to make the thing too hopeless, two positive constraints are also built into the template. One, the right context of the word holding slot is chosen to be a punctuation mark (the comma). This effectively removes the right boundary ambiguity of the word in question. Definitely, if the comma is replaced with a Chinese character string with an ambiguity level comparable with the left context "发展", the effectiveness of thus formed template will be even less predictable.

Second, the word holding slot is restricted to exactly two Chinese characters long. This effectively reduces the variety of words in question. We will report elsewhere case studies for other word holding slot lengths.

## 3.3 Data

We extracted all partial sentences matching the above template from the 4-million-character PH corpus (Guo and Lui, 1992). The complete list of all the 257 matches is in Appendix A. As the number of matches is not large, readers are recommended to have their own detailed examination. All the character pairs taking the template's word holding slot are listed in the table below. They are in the same order as in Appendix A. Duplications are not removed, since they may correspond to different context.

*Table 1: The 257 Chinese character bigrams taking the word holding slot in template "<发展><bigram><comma>". Extracted from the PH corpus. Duplication preserved.*

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 迅速 | 合作 | 较快 | 着想 | 较快 | 下去 | 速度 | 势头 | 经济 | 方向 |
| 经济 | 战略 | 规划 | 方向 | 起来 | 后劲 | 起来 | 经济 | 基金 | 养鱼 |
| 计划 | 迅速 | 指标 | 很快 | 腰果 | 道路 | 纲要 | 速度 | 目标 | 较快 |
| 潜力 | 目标 | 方向 | 畜牧 | 顺利 | 规划 | 资金 | 品种 | 生产 | 方针 |
| 壮大 | 较快 | 战略 | 同时 | 经济 | 情况 | 方面 | 更快 | 规律 | 生产 |
| 战略 | 战略 | 下去 | 很快 | 下去 | 经济 | 起来 | 时期 | 计划 | 最快 |
| 计划 | 计划 | 阶段 | 计划 | 水平 | 战略 | 项目 | 道路 | 不快 | 生产 |
| 趋势 | 后劲 | 动态 | 缓慢 | 关系 | 速度 | 服务 | 基金 | 基金 | 需要 |
| 政策 | 生产 | 生产 | 趋势 | 规划 | 较快 | 情况 | 生产 | 关系 | 农业 |
| 过多 | 过快 | 很快 | 基金 | 资金 | 规划 | 计划 | 阶段 | 迅速 | 历史 |
| 来说 | 太快 | 阶段 | 战略 | 水平 | 过程 | 实际 | 较快 | 规划 | 进程 |
| 迅速 | 纲要 | 很快 | 体育 | 服务 | 规划 | 方向 | 很快 | 生产 | 党员 |
| 较快 | 迅速 | 生产 | 中国 | 规划 | 计划 | 很快 | 情况 | 援助 | 生产 |
| 林业 | 顺利 | 任务 | 过程 | 壮大 | 阶段 | 能力 | 经济 | 道路 | 很快 |
| 较快 | 需要 | 道路 | 着眼 | 规划 | 很快 | 方向 | 经济 | 经费 | 模式 |
| 方向 | 旅游 | 变化 | 迅速 | 历史 | 农业 | 进步 | 战略 | 品种 | 缓慢 |
| 经济 | 规划 | 之中 | 迅速 | 多了 | 水平 | 重点 | 目标 | 旅游 | 情况 |
| 计划 | 公司 | 时期 | 速度 | 大计 | 规划 | 阶段 | 之路 | 之路 | 养羊 |
| 迅速 | 较快 | 迅速 | 来看 | 的县 | 上去 | 关系 | 迅速 | 战略 | 能力 |
| 任务 | 方向 | 过程 | 规划 | 壮大 | 情况 | 迅速 | 缓慢 | 起来 | 加工 |
| 需要 | 迅速 | 现状 | 方向 | 以后 | 很快 | 战略 | 经济 | 顺利 | 农业 |
| 经济 | 工作 | 减慢 | 迅速 | 态势 | 迅速 | 途径 | 重点 | 壮大 | 前景 |
| 计划 | 战略 | 需要 | 规划 | 经济 | 水平 | 速度 | 前景 | 情况 | 计划 |
| 农业 | 势头 | 探索 | 进程 | 阶段 | 绵羊 | 生产 | 蓝图 | 战略 | 条件 |
| 目标 | 中国 | 生产 | 战略 | 前途 | 关系 | 机遇 | 繁荣 | 中国 | 来说 |
| 趋势 | 战略 | 速度 | 战略 | 经济 | 基金 | 蓝图 | * | * | * |

## 3.4 Analysis

The majority listed above are legitimate dictionary words. There are only 39 bigrams not found in XianHan (1983), a famous medium size authentic Chinese dictionary with about 50,000 entries.

*Table 2: analysis of the Chinese character bigrams not in the XianHan dictionary.*

| structure | num | Unknowns | words in XianHan |
|---|---|---|---|
| adv + adj | 24 | 很快/9,较快/9,更快/1<br>最快/1,不快/1,过快/1<br>太快/1, 过多/1 | 大红 |
| prop noun | 3 | 中国/3 | 中国人民解放军 |
| verb + verb | 3 | 来说/2, 来看/1 | (而言; 说来,看来) |
| 的+ noun | 3 | 之路/2, 的县/1 | 的话 |
| verb + noun | 2 | 养鱼/1, 养羊/1 | 养兵, 养地 |
| 之 + loco | 1 | 之中/1 | 之前, 之后 |
| adj + 了 | 1 | 多了/1 | |
| verb + adj | 1 | 减慢/1 | 减低,减轻,减弱,减少 |

Among the 39 non-dictionary bigrams, the dominant portion are those following the compounding pattern: ***"adv. + adj."***, where both the adverb and the adjective are a single Chinese character. There are totally 7 distinct types (很快(very fast)/9, 较快(relatively fast)/9, 更快(faster)/1, 最快(fastest)/1, 不快(not fast)/1, 过快(too fast)/1, 太快(overly fast)/1, 过多(too many)/1) accumulatively occurred 24 times. Note, ***"adv. + adj."*** is a valid Chinese word formation pattern (the so-called ***状中结构, modifier-predicate compounding***). There do exist in XianHan words like "大红 (bright red)" following this pattern. Moreover, all these bigrams are with high usage frequency and expressive mutual information score. That is, they are all "tight-in-combination and stable-in-use". Because

Chinese words, phrases, and sentences are formed under the same principle (Zhu, 1985), there is essentially no clear boundary for words, phrases and even sentences. Whether or not they are words might largely depend on the taste of individual lexicographers. Nevertheless, they are all legitimate and self-contained syntactic and/or semantic entities.

The rest 14 occurrences represents seven different types. Proper nouns such as country names like "中国 (China)" above are normally not collected in the main body of a dictionary. Rather, they are conventionally compiled as supplementary dictionary appendices. What we want to emphasize here is that such sub-language entities emerge themselves naturally in context.

"来说" and "来看" are not recorded in XianHan, but they are nevertheless tightly bounded and frequently used. In addition, they have quite unique characteristics in usage and peculiar (syntactic) meaning.

"的县" and "之路" are not in XianHan. But "的话" does. Similarly, XianHan has "之后" and "之前" but no "之中". Both "养羊" and "养鱼" are not in XianHan, but "养兵" and "养地" are in it. XianHan has "减低", "减轻", "减弱" and "减少", but no "减慢".

### 3.5 Conclusion

Frankly, I would hesitate to quantify the precision and/or recall achieved from the particular template above, as there exist significant inter-judge variances on whether or not what listed above are words. Instead, I make relevant data available in full and suggest readers to do their own calculation. What I hope to accomplish in this section is to convince readers the following two observations.

(1) Word/chunk identification/extraction by template matching is effective. At least this is true for the particular non-trivial template of "<发展><bigram><comma>". This observation is important as it provides us with an empirical support of our context-centered template matching word identification scheme.

(2) The effectiveness of a template in word identification is quantitatively measurable with respect to given corpus and dictionary. This observation is important as it implies the learnability or trainability of word identification templates. That is, we may have automatic ways for template preparation.

## 4 MATHEMATICAL MODELING

In this section, we will put the context-centered thinking into a broad world. Through formal mathematical modeling, we will achieve deep understanding on both the word identification problem and its different problem solving strategies. Entity modeling and context modeling are to be introduced and contrasted.

### 4.1 The World

Given context as a word holding template and entity taking up the word holding slot of the template, the question here is: whether or not the entity is a word. Assume there are only two clear-cut answers: "***yes***" or "***no***". Then, in the language of ***probability***, we have created a tiny world with three citizens or ***random variables***: the ***word holding template T*** taking set of mutually independent templates as the universe, the ***word holding slot S*** which could be filled up with certain type of entities, and the ***answer A*** to the question taking either "***yes***" or "***no***". Moreover, the joint probability

*(1)* $Pr(A, S, T)$

gives us the precise and complete description of the three-random-variable world. That is, having the joint probability (1) known, any question about the world could be answered.

### 4.2 Word Identification Modeling

In particular, suppose word holding template $T=t$ match a natural language sentence extracted from a corpus, and $S=s$ be the sentence fragment taking the word holding slot. Then, it has been well established (Duda and Hart, 1973) that, on average, the ***minimum error decision*** is, $S=s$ is a word in sentence if and only if there holds

(2) $Pr(A=yes, S=s, T=t) > Pr(A=no, S=s, T=t).$

Assume

(3) $Pr(T,S|A) = Pr(T|A)Pr(S|A),$

and denote

(4) $Lt = ln\ Pr(A=yes|T=t) / Pr(A=no|T=t),$

(5) $Ls = ln\ Pr(A=yes|S=s) / Pr(A=no|S=s),$

(6) $La = ln\ Pr(A=yes) / Pr(A=no),$

the decision rule (2) could be rewritten as

(7) $Lt + Ls > La.$

That is, under the assumption given in formula (3), to have minimum error solution to the word identification problem is equivalent to calculate the ***context likelihood Lt***, the ***entity likelihood Ls*** and the ***solution likelihood La***, and to make decision with rule (7). This is in turn equivalent to estimate the ***context probability Pr(A\|T)***, the ***entity probability Pr(A\|S)*** and the ***solution probability Pr(A).***

## 4.3 Entity Modeling

The task of entity modeling is to estimate for any possible entity $S=s$ the entity likelihood defined in formula (5):

(5) $Lt = ln\ Pr(A=yes|S=s) / Pr(A=no|S=s).$

As we elaborated before, depending on system configuration, an entity could be a plain character string, a string of simple words, or some type of character, word, and part-of-speech combinations. We use the term ***entity*** for generality.

Note that, since we have assumed only two possible answers, there holds

(8) $Pr(A=yes|S=s) + Pr(A=no|S=s) = 1.$

Then, the entity likelihood could be equivalently written as

(9) $Lt = ln\ Pr(A=yes|S=s) / (1 - Pr(A=yes|S=s)).$

Hence, the heart of entity modeling is in modeling the probability $Pr(A=yes|S=s)$ for any possible entity $s$. In essence, the probability $Pr(A=yes|S=s)$ is expressing the fitness of entity $S=s$ as a word out of context.

Entity modeling is nothing but the theme in both sub-language oriented and entity-centered word identification research. Numerous modeling approaches have been proposed in literature. For example, in Lee, Lee, and Chen (1994), the probability of character trigram $C_1C_2C_3$ used as a Chinese name is estimated as the product of $Pr(C_1|surname)$, the probability of the first character $C_1$ as a surname, $Pr(C_2|middlename)$, the probability of the second character $C_2$ as a middle name, and $Pr(C_3|givenname)$, the probability of the third character $C_3$ as a given name:

(10) $Pr(A=yes|S=C_1C_2C_3)$
$= Pr(C_1|surname)\ Pr(C_2|middlename)\ Pr(C_3|givenname).$

The ngram-based approach by Fung and Wu (1994) is in fact implicitly modeling $Pr(A=yes|S=ngram)$ through a statistical decision procedure, while their linguistic filters (Wu and Fung, 1994) are implicitly modeling $Pr(A=no|S=ngram)$ through a set of linguistic selection rules.

## 4.4 Solution Modeling

The task of solution modeling is to estimate the solution likelihood defined in formula (6):

(6) $La = ln\ Pr(A=yes) / Pr(A=no).$

Similar to what we did for entity modeling, only the probability $Pr(A=yes)$ is to be estimated. Theoretically, this could be done directly from a corpus by counting the cases where positive and negative decisions are made.

In practice, however, the likelihood of $La=lnPr(A=yes)/Pr(A=no)$ is better obtained from user assignment than from corpus training. Similar to what we did in Chinese spell checking and proofreading (Guo, 1994), the solution likelihood could be used as a control variable for recall/precision balancing. Different solution likelihood settings will result in different recall and precision rates. You may be able to get high precision at the cost of low recall, or to go the other way around. We found this is a practical mechanism for fulfilling different preferences within the single system.

## 4.5 Context Modeling

The task of context modeling is to estimate from any possible word holding template $T=t$ the context likelihood defined in formula (4):

$$(4) \qquad Lt = ln\ Pr(A=yes|T=t) / Pr(A=no|T=t).$$

Similar to the discussion above, this likelihood could also be written as

$$(11) \qquad Lt = ln\ Pr(A=yes|T=t) / (1 - Pr(A=yes|T=t)).$$

The core here is to estimate $Pr(A=yes|T=t)$, the probability of entity taking the word holding slot of the word holding template $T=t$ a valid word. Given a tokenized corpus, its ***Maximum Likelihood Estimation*** (*MLE*) is:

$$(12) \qquad Pr(A=yes|T=t) = n/N$$

where $N$ = "the number of times template $T=t$ matches a corpus sentence" and $n$ = "the number of times the entity in the matched template slot is a valid word".

For example, in the case study above, there is $N=257$. If we treat only those in XianHan (1983) as valid word, there is $n=219$. Thus, there are:

$$Pr(A=yes|T="<发展><bigram><comma>")=219/257=0.85,$$

$$Lt = ln\ 219/38=1.75.$$

If only the three cases under the pattern ***"的 + noun"*** are considered non-words, there are:

$$Pr(A=yes|T="<发展><bigram><comma>")=254/257=0.99,$$

$$Lt = ln\ 254/3=4.44.$$

## 4.6 Template-Slot Independence Assumption

It must be pointed out that, to reach the word identification model (7), we explicitly made the following assumption in formula (3):

$$(3) \qquad Pr(T,S|A) = Pr(T|A)Pr(S|A).$$

This assumption could be read as that, with respect to any specific answer $A=a$, the word holding template $T$ and the word holding slot $S$ are probabilistically independent. In this paper, this is referred to as ***Template-Slot Independence Assumption***.

This assumption does not hold for templates such as parts of collocational patterns. This assumption is adopted for two reasons. First, we are not aware of any practically computable modeling approach which could take into account all the three random variables simultaneously. We have to decompose the three variable world into several one or two variable subworlds. That is, we have to accept some compromise.

Second, tight-bounding collocations are relatively rare in real text. For an English corpus of 10 million words, according to Smadja (1993), only about 8,000 collocations are reliable, a neglectable portion among millions of other bigrams and ngrams. Moreover, as words which could take up slots in such collocational templates are highly constrained, such collocations, even if chosen as word holding templates, are not productive.

In short, the assumption we adopt is necessary for efficient computational modeling and feasible for productive practical application.

### 4.7 Short Summary

According to the word identification model given in formula (7), to solve the problem unbiasly, both entity modeling and context modeling are necessary. In literature, however, the emphasis has almost exclusively been put on entity modeling and it is hardly possible to find a proposal with the context model explicitly incorporated.

In contrast, by exploring the word identification problem in such a systematic way, we make it explicit the importance of context modeling, and proposed a concrete context-modeling approach.

The use of context information in such an explicit and systematic way is the fundamental difference between the approach proposed in this paper and the rest in literature.

## 5 ALGORITHM

In this section, we will first go to the extreme by trying to construct lexicon and perform tokenization ***without dictionary.*** Such an algorithm is proposed. Some notes, variations and improvements are then given to contrast the key idea and to make the thing realistic.

### 5.1 Start from Empty

In Section 3, we presented in detail a case study which gives us strong confidence on context-centered word identification and lexicon construction. Note, in that case study, we did not do any preprocessing. In particular, we did not do text word tokenization. Moreover, we did not use any dictionary in word identification. The dictionary XianHan (1983) is employed only for human evaluation of the effectiveness of the scheme. Explicitly, we are building up lexicon from scratch and doing text tokenization without lexicon. This section is on the algorithm proper.

### 5.2 Baseline Algorithm

The baseline algorithm has four phases:

#### *5.2.1 Initialization (n=1)*

- Put all none Chinese character symbols in the working character set such as GB, Big5 or UNICODE into the lexicon. Punctuation marks, numerical digits and Roman alphabets are part of these symbols.
- Put all function words, such as prepositions, pronouns, particles and classifiers, into the lexicon.
- Put a few special Chinese characters such as "是" and "有" into the lexicon.
- Put a special sentence begin symbol and a special sentence end symbol into the lexicon.
- Prepare a large text corpus by appending the sentence begin and sentence end symbols at the two ends of every sentence in corpus.
- Set a maximum word length.

#### *5.2.2 Bootstrapping (n=2)*

- For each word $W_1$ and $W_2$ in the above initialized lexicon, form word holding template $T(W_1,W_2)$ = ***<W1><Chinese character bigram><W2>,*** where the word ***W1*** and ***W2*** are the left and right context respectively, and the ***Chinese character bigram*** takes the word holding slot.
- For each thus formed word holding template and for each sentence in corpus, check to see if there is a match.
- If yes, put the Chinese character bigram found in the matched Chinese sentence and taking the word holding slot of the word holding template into lexicon.

### *5.2.3 Iteration (n>2)*

Suppose lexicon up to word length n-1, n>2, has been constructed. Now, to augment the lexicon with words of length n,

- For each word $W_1$ and $W_2$ in the previously constructed lexicon, form word holding template $T(W_1,W_2)$ = ***<W1><Chinese character ngram><W2>***, where the word *W1* and *W2* are the left and right context respectively, and the ***Chinese character ngram*** takes the word holding slot.
- For each thus formed word holding template, evaluate its context likelihood *Lt* for word holding slot of length $m<n$ Chinese characters. This is done by (1) matching the modified (that is, the requirement for the word holding slot has been relaxed to allow any Chinese $m<n$ grams) word holding template against the whole corpus, (2) counting the number of m-grams which are taking the word holding slot and are or are not words in the latest lexicon, and (3) calculating the log ratio according to the formula $Lt = ln\ Pr(A=yes|T=t) / Pr(A=no|T=t)$ introduced in the last section.
- If the context likelihood is better than a preset threshold (the solution likelihood), the template is confirmed.
- For each thus confirmed word holding template and for each sentence in corpus, check to see if there is a match.
- If yes, put the Chinese character ngram, which is found in the matched Chinese sentence and takes the word holding slot of the confirmed word holding template, into lexicon.

Repeat the above procedure successively for each n, until it reaches the preset maximum word length limit.

### *5.2.4 Completion (n=1)*

- Add all single Chinese characters into the lexicon.

## 5.3 Notes

First, in Phase 2, only bigrams are collected, and bigrams involving none Chinese character symbols are excluded.

Second, compared with Phase 2, template evaluation and confirmation are added to Phase 3. Operations in this phase will be executed consecutively for n=3 up to the preset maximum word length.

Third, the last phase, Phase 4, is to make the lexicon complete. As defined in (Guo, 1996), a lexicon is complete if all words in open text are in the lexicon. Operationally, this is proven to be equivalent to have all characters in the working character set be included in the lexicon (Guo, 1996).

## 5.4 Improvements and Variations

Countless improvements and variations of the baseline algorithm are possible. To keep the description brief, only a few are discussed below.

### *5.4.1 "Top Star" Templates*

In the baseline algorithm above, both the left and right template context are single word long. This could be updated to allow multi-word context for higher reliability. Most rare templates are better to be discarded in production, since their word predicting power could not be faithfully estimated. A practical way is to generate various context length templates but only keep those appearing at least several dozen times in the corpus. This trick will both boost the lexicon's quality and the algorithm's computational efficiency. The efficiency comes from the fact that, only high frequency entries in the lexicon need to be considered in template preparation, evaluation and matching, yet these high frequency entries are always a tiny portion of the whole lexicon. Moreover, the frequent appearance of a template in corpus makes the estimation of its context likelihood operationally reliable, and thus reduce the noise level.

Methodologically, this is essentially ***to let a few "top star" context templates call back the "silent majority" slot words.***

### *5.4.2 Resource*

In what we did above, there is neither lexicon nor lexicon-based tokenizer employed. However, the two resources are not too difficult to obtain. Suppose we already have a machine readable high quality dictionary and a lexicon-based tokenizer such as the simplest longest (maximum) matching segmentor. Then, we could start by using that dictionary as initial lexicon. Moreover, before each iteration begin, we could tokenize the corpus according to the current available lexicon. This will help us both in reducing errors caused by cross word boundary template matching and in enhancing computational efficiency through precompiling corpus ngram statistics. Furthermore, part-of-speech tagger could also be incorporated.

### *5.4.3 Linear Pattern Matching*

It is also worth noting here the ***Aho-Corasick*** pattern matching algorithm (Aho and Corasick, 1975, Aho, 1990) and its derivation (Pinter, 1985) which allows the inclusion of don't-care symbols in patterns. By organizing lexicon and/or templates in ***trie*** (Knuth, 1973) structure, the pattern matching can be implemented in time linearly proportional to the corpus size and independent of the size of lexicon and/or template. However, detailed discussions on efficiently applying these algorithms to lexicon-based tokenizations are out of the scope of this paper. Interested readers are referred to (Guo, 1996).

## 6 CONCLUSIONS

In this paper, we have established the fresh context-centered Chinese lexicon construction methodology which is first introduced with an intuitive explanation and an informative case study, and then generalized with formal mathematical modeling and algorithm development. Due to space limit, experimental evaluations will be reported elsewhere.

The primary advantage of the context-centered lexicon construction approach is its unmatchable productivity of recalling the "silent majority". The algorithm we proposed could extract almost all words and phrases in text, even if they only appear once or twice. Moreover, as our case study depicted, the precision could also be very good.

Under the guidance of our mathematical word identification model, several possible improvements are also highlighted in this paper. It is understood that we have restricted ourselves to be within the framework of traditional syntactic analysis. What we believe to be worth further pursuing are the following two aspects:

- ***Chunks***: In English, a chunk is essentially a "non-recursive core of an intra-clausal constituent" (Abney, 1991, Abney, 1996). To make it applicable to Chinese, some adaptations are required. The principle, however, is nevertheless there. Only after introducing the concept of chunks, we can have a better understanding of our task. In Chinese, lexicon construction is nothing but chunk collection. It is practically not critical to argue the definition of words. Rather, if an ngram is a chunk, we collect it.
- ***Semantics***: The more effective way of enhancing the quality of a lexicon is by introducing semantic databases such as the multilingual ***SenseWeb*** (Dong and Guo, 1996) or the English ***WordNet*** (Miller, 1990). Normally we do not work from scratch but augment an already validated lexicon. Then, with SenseWeb or WordNet, we could check the similarity between what we are extracting from corpus and what are already in the lexicon.

## ACKNOWLEDGEMENTS

Great thanks to Prof. Dong Zhen Dong and Dr. Robert Luk for insightful discussion, genius help and paper critiquing. Nurdini helped me a lot in improving the English writing.

## REFERENCES

Abney, Steven (1991), ***Parsing by Chunks***, In Robert Berwick, Steven Abney and Carol Tenny (eds.), Principle-Based Parsing, Kluwer Academic Publishers, Dordrecht.

Abney, Steven (1996), ***Chunk Stylebook***, On-line document, http://www.sfs.nphil.uni-tuebingen.de/~abney/96i.ps.gz.

Aho, A. V., (1990), ***Algorithms for Finding Patterns in Strings***, in: (ven Leeuwen, J., eds.) Handbook of Theoretical Computer Science, Volume A, Algorithms and Complexity, Chapter 5, pp. 273-278, The MIT Press.

Aho, A. V., and Corasick, M. J., (1975), ***Efficient String Matching: An Aid to Bibliographic Search***, Communications of ACM, 18 (6), pp. 333-340.

Black, Ezra, Roger Garside, and Geoffery Leech, (1993), ***Statistically-Driven Computer Grammars of English: The IBM/Lancaster Approach***, Amsterdam: Rodopi Publishers.

Chang, J. S., Chen, S. D., Ker, S. J., Chan, Y. and Liu J. S., (1994), ***A Multi-Corpus Approach to Recognition of Proper Names in Chinese Texts***, Computer Processing of Chinese and Oriental Languages, Vol. 8, No. 1, page 73-86.

Chen, Keh-Jiann, and Shing-Huan Liu, (1992), ***Word Identification for Mandarin Chinese Sentences***, In Proceedings COLING-92, Nantes, pp. 101-107.

Chiang, Tung-Hui, Jing-Shin Chang, Ming-Yu Lin, and Keh-Yih Su, (1992), ***Statistical Models for Word Segmentation and Unknown Word Resolution***, In Proceedings of ROCLING-92, pp. 121-146.

Church, Kenneth. W., and William A. Gale, (1991), ***A Comparison of the Enhanced Good-Turing and Deleted Estimation Methods for Estimating Probabilities of English Bigrams***, Computer Speech and Language, Vol. 5, No. 1, page 19-54.

Church, Kenneth. W., and P. Hanks, (1990), ***Word Association Norms, Mutual Information and Lexicography***, Computational Linguistics, Vol. 16, No. 1, pp. 22-29.

Dong, Zhen Dong and Jin Guo, (1996), ***Construction and Application of SenseWeb -- a Multilingual Semantic Lexicoal Database***, A Tutorial at the 1996 International Conference on Chinese Computing (ICCC96), Singapore.

Duda, R. O., and P. E. Hart, (1973), ***Pattern Classification and Scene Analysis***, New York, John Wiley & Sons.

GB-13715, (1993), ***Contemporary Chinese Language Word Segmentation Specification for Information Processing***, PRC National Standard, China National Standard Bureau.

Guo, Jin, and Ho Chung Lui, (1992), ***PH: a Chinese Corpus for Pinyin-Hanzi Transcription***, Technical Report TR93-112-0, Institute of Systems Science, National University of Singapore.

Guo, Jin, (1993), ***Statistical Language Modeling and Some Experimental Results on Chinese Syllables to Words Transcription***, Journal of Chinese Information Processing, Vol. 7, No. 1, pp. 18-27.

Guo, Jin, (1994), ***Automatic Chinese Spelling Checking***, A Tutorial at the 1994 International Conference on Chinese Computing (ICCC94), Singapore, Part I, 31 pages, Part II, 44 pages.

Guo, Jin, (1996), ***An Efficient and Complete Algorithm for Unambiguous Word Boundary Identification***, (to be submitted), available on-line: http://sunzi.iss.nus.sg:1996/guojin/papers/acbci.ps.gz.

Fung, Pascale, and Dekai Wu, (1994), ***Statistical Augmentation of a Chinese Machine-Readable Dictionary***, In Proceedings of the Second Annual Workshop on Very Large Corpora (WVLC-2), Kyoto.

Harman, Donna K. (eds.) (1995), ***Overview of the Third Text REtrieval Conference (TREC-3),*** NIST Special Publication, Washington, DC: US Government Printing Office.

Harman, Donna K. (eds.) (1996), ***Overview of the Fourth Text REtrieval Conference (TREC-4),*** NIST Special Publication, Washington, DC: US Government Printing Office.

Hu, Yu Shu, (1987), ***Xiandai Hanyu (Modern Chinese),*** Shanghai Education Press.

Knuth, D. E., (1973), ***Fundamental Algorithms***, second edition, The Art of Programming, Vol. 1. Addison-Wesley, Reading, Mass., pp. 487-499.

Kucera, H., and Francis, W. N., (1967), ***Computational Analysis of Present-Day American English***, Providence, Brown University Press.

Lee, C, Y. Lee, and H. Chen, (1994), ***Research of the Identification of Names in Chinese Text***, In Proceedings of ROCLING VII, pp. 203-222.

Lin, Ming-Yu, Tung-Hui Chiang, and Keh-Yih Su, (1993), ***A Preliminary Study on Unknown Word Problem in Chinese Word Segmentation***, In Proceedings of ROCLING VI, pp. 119-141.

Lua, K. T., (1995), ***Experiments on the Use of Bigram Mutual Information in Chinese Natural Language Processing***, In Proceedings of the 1995 International Conference on Computer Processing of Oriental Languages (ICCPOL-95), Hawaii, pp. 306-313.

Luk, Robert W. P., (1995), ***Automatic Tokenization of Chinese Text Driven by a Lexicographic Index with Linguistic Pattern Filtering***, In Proceedings of the 1995 International Conference on Computer Processing of Oriental Languages (ICCPOL-95), Hawaii, pp. 217-224.

Luk, Robert, W. P., (1996), Personal Communications.

Markus, M., M. A. Marcinkiewicz, and B. Santorini, (1993), ***Building a Large Annotated Corpus of English: The Penn Treebank***, Computational Linguistics, Vol. 19, No. 2, page 313-329.

Miller, George A., (1990), ***WordNet: An On-line Lexical Database***, International Journal of Lexicography, Vol. 3, No. 4, page 235-312.

Mo, Ruo-ping, Yao-Jung Yang, Keh-Jiann Chen, and Chu-Ren Huang, (1991), ***Determinative-Measure Compounds in Mandarin Chinese: Formation Rules and Parser Implementation***, In Proceedings of ROCLING-IV, pp. 111-134.

Nie, J-Y., Jin W-Y., and Hannan, M-L., (1994), ***A Hybrid Approach to Unknown Word Detection and Segmentation of Chinese***, In Proceedings of International Conference on Chinese Computing 1994 (ICCC-94), Singapore, page 326-335.

Pinter, R. Y., (1985), ***Efficient String Matching with Don't-Care Patterns***, in: A. Apostolico and Z. Galil, (eds.), Computational Algorithms on Words, Springer Berlin.

Smadja, Frank, (1993), ***Retrieving Collocations from text: Xtract***, Computational Linguistics, Vol. 19, No. 1, pp. 143-177.

Song, Rou, Chaojie Qiu, Longgeng Ouyang, Lubing Xu, and Xin Wang, (1996), ***Bi-Orderly-Neighborship and Its Application to Chinese Word Segmentation and Proofreading***, In Proceedings of the 1996 International Conference on Chinese Computing (ICCC96), Singapore, pp. 428-433.

Sproat, Richard, and Shih, Chilin, (1990), ***A Statistical Method for Finding Word Boundaries in Chinese Text***, Computer Processing of Chinese and Oriental Languages, Vol. 4, No. 4, page 336-349.

Sproat, Richard, (1994), ***English Noun-Phrase Accent Prediction for Text-to-Speech***, Computer Speech and Language, 1994, No. 8, pp. 79-94.

Sun, Maosong, and Changning Huang, (1996), ***Word Segmentation and Part-of-Speech Tagging for Unrestricted Chinese Texts***, A Tutorial at the 1996 International Conference on Chinese Computing (ICCC96), Singapore.

Wu, Dekai, and Pascale Fung, (1994), ***Improving Chinese Tokenization with Linguistic Filters on Statistical Acquisition***, In Proceedings of the 4th Conference on Applied Natural Language Processing (ANLP-94), Stuttgart, pp. 180-181.

Wu, H-J., Jin Guo, Ho Chung Lui., and Hwee Boon Low., (1994), ***Corpus-Based Speech and Language Research in the Institute of Systems Science***, In Proceedings of International Symposium on Speech, Image Processing and Neural Networks (ISSIPNN-94), Hong Kong, pp. 142-145.

XianHan (1983), ***现代汉语词典 (Modern Chinese Dictionary)***, 2nd edition, Commercial Press, Beijing.

Yang, Jin, and Laurie Gerber, (1996), ***SYSTRAN Chinese-English Machine Translation System***, In Proceedings of the 1996 International Conference on Chinese Computing (ICCC96), Singapore, pp. 205-210.

Yuan, Baosheng, Yuqin Gao, Hisao-Wuen Hon, Jean-Luc Lebrun, Zhiwei Lin, Gareth Loudon, and Xi Han, (1996), ***Chinese Dictation Kit: A Very Large Vocabulary Mandarin Speech Input System***, In Proceedings of the 1996 International Conference on Chinese Computing (ICCC96), Singapore, pp. 1-4.

Zheng, J. and K. Liu, (1993), ***Approach of Processing Tactics on the Names and Surnames in Chinese Automatic Segmenting System***, In Chen and Yuan (eds.) Computational Linguistics: Research and Applications, Beijing Linguistics Institute Publisher, Beijing, pp. 139-143.

Zhu, Dexi, (1985), ***语法答问 (Questions and Answers on Chinese Grammar)***, Commercial Press, Beijing.

## APPENDIX

Listed below are the 257 partial sentences in the PH corpus matching template "<发展><bigram><comma>". See section 3 for analysis.

1.制盐业和有色金属产业发展迅速,
2.苏联外长谢瓦尔德纳泽和１９日抵苏访问的民主德国外长菲舍尔在２０日会谈中主张在华约范围内进一步发展合作,
3.全区勤工俭学近几年发展较快,
4.鞍钢党政领导从企业的长期发展着想,
5.技术出口发展较快,
6.稳定地发展下去,
7.在困难的条件下保持了正常的发展速度,
8.我国地方铁路建设呈现持续发展势头,
9.努力发展经济,
10.关于乡镇企业今后的发展方向,
11.它对发展经济,
12.实现分三步走的经济发展战略,
13.二是制定了发展规划,
14.稳定面积攻单产的发展方向,
15.把生产力发展起来,
16.增强农业发展后劲,
17.羽绒厂等加工企业迅速发展起来,
18.向全县人民发出了“发展经济,
19.并尽快建立竹类科技发展基金,
20.现在全区已有２０多万劳力从事专业发展养鱼,
21.在研究制定经济和社会发展计划,
22.乡镇企业起步晚但发展迅速,
23.强调以农业特别是粮食发展指标,
24.农村经济发展很快,
25.在沿海发展腰果,
26.找到并坚持了符合本国国情的发展道路,
27.省政府制订了农田水利发展纲要,
28.这一惊人的发展速度,
29.生态发展目标,
30.卷烟工业发展较快,
31.又有发展潜力,
32.对围绕产业发展目标,
33.低残留的发展方向,
34.发展畜牧,
35.两国在各个领域的友好合作关系发展顺利,
36.孟加拉国政府制订了长期发展规划,
37.由政府向全国足球协会提供发展资金,
38.发展品种,
39.帮助群众发展生产,
40.抓落实”的发展方针,
41.指出中国的振兴最终要靠高科技和新技术产业的发展壮大,
42.去年捷同非社会主义国家的易货贸易发展较快,
43.江西省各级党政部门自觉围绕这一发展战略,
44.在畜牧业稳步发展同时,
45.会议呼吁发达国家帮助最不发达国家更快发展经济,

46.他这次访泰的目的是研究柬埔寨局势发展情况，
47.找到最佳发展方面，
48.发展更快，
49.不断揭示各种灾害的成因和发展规律，
50.不少地方应用这种方式发展生产，
51.各国政府有责任根据本国国情制定发展战略，
52.制定出切合本国实际的发展战略，
53.如此发展下去，
54.近年来亚太地区的经济发展很快，
55.如果这种局势发展下去，
56.南亚各国才能发展经济，
57.我国海洋石油工业是在改革开放中发展起来，
58.今年我国旅游业进人了恢复和发展时期，
59.批准１９８９年国民经济和社会发展计划执行情况和１９９０年国民经济和社会发展计划，
60.步人了历史上又一个发展最快，
61.批准１９８９年国民经济和社会发展计划执行情况和１９９０年国民经济和社会发展计划，
62.批准１９８９年国民经济和社会发展计划执行情况和１９９０年国民经济和社会发展计划，
63.进人了一个新的发展阶段，
64.我们制定方针政策和经济社会发展计划，
65.多数民族自治地区将会逐步接近或赶上当时全国的中等发展水平，
66.根据国家的产业政策和发展战略，
67.农林牧方面的发展项目，
68.使匈牙利迈向现代化的发展道路，
69.目前铁路发展不快，
70.发展生产，
71.对我国横向经济联合的发展趋势，
72.增强企业发展后劲，
73.密切注视国际造船业的发展动态，
74.欧洲企业减少在非洲的投资主要是因为非洲市场发展缓慢，
75.各地区及有关国际组织保持和发展关系，
76.工业生产保持一定的发展速度，
77.新课程的设置与发展服务，
78.建立农业发展基金，
79.海安县利用筹集的约２０００万元农业发展基金，
80.又结合香港的实际情况和发展需要，
81.他们希望政府从速制订中长期发展政策，
82.牲畜和农用机具等物资以发展生产，
83.企业承包经营责任制无论是对发展生产，
84.充分了解和科学地预测世界高技术发展趋势，
85.大多数省市都已制定了特殊教育的发展规划，
86.西藏自治区社会福利事业近年来发展较快，
87.双方讨论了巴勒斯坦问题的最近发展情况，
88.双方都要求通过紧密的协作发展生产，
89.特别是伊朗发展关系，
90.进一步发展农业，
91.宾馆饭店发展过多，
92.近几年棉纺能力发展过快，
93.蔡塘村发展很快，
94.山东省各级财政部门还将征集的４亿多元的农业发展基金，
95.这个省还注意积累渔业发展资金，
96.成立了这个旨在研讨产业政策布局和发展规划，
97.根据中国的发展计划，
98.新疆维吾尔自治区民族团结进人一个新的发展阶段，
99.保险业务发展迅速，
100.回顾我们党的新闻事业的发展历史，
101.就新闻战线自身的发展来说，
102.由于近年来新闻队伍发展太快，
103.中国出版业跨人了全新的发展阶段，
104.必须服从国家的总体发展战略，
105.不能脱离我国的经济文化整体发展水平，
106.回顾总结党的十一届三中全会以来的历史发展过程，
107.紧密围绕天津教育发展实际，
108.港口等单位运输１０吨集装箱的业务发展较快，
109.制定出一个发展规划，
110.发达国家限制纺织品进口阻碍了发展中国家的发展进程，
111.我国戏曲电视剧近年发展迅速，
112.科学技术长期合作发展纲要，
113.“近年来首钢的事业发展很快，
114.“发展体育，
115.为各项事业的发展服务，
116.市政府就制定了“菜篮子”工程５年发展规划，
117.麦棉套种是发展方向，
118.国外某个领域的技术发展很快，
119.尽快恢复和发展生产，
120.大冶钢厂党委重视在生产一线发展党员，
121.除了江南地区近几年经济发展较快，
122.由于发展迅速，
123.努力发展生产，
124.才能发展中国，
125.我国高技术研究发展规划，

126.为了今后更好地实施“８６３”高技术研究发展计划，
127.河南省小火电机组近年来发展很快，
128.双方回顾了１０年来两国教育学术交流的发展情况，
129.即把它们国民生产总值的百分之零点七提供为官方发展援助，
130.继续努力发展生产，
131.龙泉市重视依靠科技发展林业，
132.“如果事态发展顺利，
133.按照全会提出的国民经济和社会发展任务，
134.从单项服务到系列化服务的发展过程，
135.集体经济越是发展壮大，
136.这标志着我国社会主义现代化建设进入了一个新的发展阶段，
137.对增强贫困地区经济发展能力，
138.波兰现在正进行改革和发展经济，
139.走什么样的发展道路，
140.蒙中两国关系在各个领域里发展很快，
141.这个地区东部发展较快，
142.产业结构调整及外向型企业发展需要，
143.选择什么发展道路，
144.二是从贫困地区的长期发展着眼，
145.必须把扶贫开发列入本地区的国民经济和社会发展规划，
146.也不会发展很快，
147.要在对社会实际情况进行深入分析和研究的基础上确定发展方向，
148.不仅要求重视发展经济，
149.国务院有关部门还拨给西藏一些专项教育发展经费，
150.科研机构要积极探索多种形式的管理和发展模式，
151.选择什么发展方向，
152.发展旅游，
153.随着国际形势和匈中两国情况的发展变化，
154.我国橡胶工业发展迅速，
155.他在谈话中回顾了匈中关系的发展历史，
156.就是要充分运用一切科技成果发展农业，
157.共同为中华妇女事业的发展进步，
158.有必要重新审视上海的发展战略，
159.发展品种，
160.因而奶业的发展缓慢，
161.发展经济，
162.十年改变贫困面貌的发展规划，
163.竭力把科学技术物化活化在经济与社会发展之中，
164.西藏自治区卫生事业发展迅速，
165.个体和私营经济不是发展多了，
166.达到小康发展水平，
167.钢铁等基础工业是中国“十年规划”和“八五计划”中的发展重点，
168.既要考虑有利于解决经济发展中的深层次问题和实现中长期的发展目标，
169.加之具有长期贸易和旅游历史的一些其他国家仍在大力发展旅游，
170.阿里维勃沃一行应国务院特区办邀请前来了解中国经济特区的发展情况，
171.农业部将尽快修订颁发“八五”农机化发展计划，
172.由内地大学毕业生为主组建的新疆唐布拉贫困与发展公司，
173.检察技术工作目前正处于一个重要的发展时期，
174.之发展速度，
175.一些著名科技专家最近在京聚会研讨我国航天事业发展大计，
176.这个省各级税务部门围绕本地经济发展规划，
177.财贸工会进入了新的发展阶段，
178.贵州虹山轴承厂走外向型企业发展之路，
179.贵州虹山轴承厂走外向型企业发展之路，
180.村里发展养羊，
181.发展迅速，
182.体育事业发展较快，
183.以乡镇工业为主体的农村非农产业发展迅速，
184.因为从香港长远的经济发展来看，
185.就是涌现了一批经济持续稳定发展的县，
186.一个村的经济能不能发展上去，
187.同日本自民党和社会党以及日本其他政党扩大与发展关系，
188.但女子足球在朝鲜却发展迅速，
189.李铁锤制定了以技术进步为主导的发展战略，
190.初步形成了自我调节和自我发展能力，
191.按照《建议》提出的国民经济和社会发展任务，
192.它进一步明确了农村的基本经济政策和发展方向，
193.从单项服务到系列化服务的发展过程，
194.农村工作开创新局面的发展规划，
195.集体经济越是发展壮大，
196.双方相互介绍了各自国家当前科技和经济发展情况，
197.广西对外保险事业发展迅速，
198.多年来发展缓慢，
199.使服务体系尽快地发展起来，

200.有些产品可以采取城乡联合或协作的办法发展加工，
201.适应旅游业发展需要，
202.近年来我国农业机械化事业发展迅速，
203.国家科委根据世界科学技术发展现状，
204.这一技术已被公认为通信网的发展方向，
205.村办工业发展以后，
206.珠海十年来发展很快，
207.共同探讨电信事业的发展战略，
208.为进一步帮助边疆地区发展经济，
209.台资在这里发展顺利，
210.重视发展农业，
211.为调动各个方面的积极因素发展经济，
212.以确保妇女更加踊跃地参与国家的社会和经济发展工作，
213.生产发展减慢，
214.改革开放以来通讯事业发展迅速，
215.为各级政府了解社会经济发展态势，
216.在短期内发展迅速，
217.发展途径，
218.确定发展重点，
219.保护现有企业的生产力并使之不断发展壮大，
220.两国牢固的关系有着广阔的发展前景，
221.被列人市国民经济发展计划，
222.从北往南推进的发展战略，
223.中国将根据经济发展需要，
224.他准备到国家民委汇报全州发展规划，
225.发展经济，
226.是适应我国生产力发展水平，
227.今后１０年保持百分之六的年均发展速度，
228.这次会关系大陆发展前景，
229.双方相互介绍了各自国家当前科技和经济发展情况，
230.同时港府正在执行一项研究和发展计划，
231.依靠科技进步发展农业，
232.显示出蒸蒸日上的发展势头，
233.十年的发展探索，
234.符合中国的国情和历史发展进程，
235.李鹏的报告清楚地表现出中国第三代领导把９０年代视为标志着中国进人一个新的发展阶段，
236.如果在南方建立人工草地１亿多亩发展绵羊，
237.贵阳市劳动局还想方设法帮助集体企业发展生产，
238.“李鹏总理的报告勾画出了今后十年的发展蓝图，
239.制订国家长期科技发展战略，
240.也是改善妇女参与社会发展条件，
241.实现教育发展目标，
242.只有社会主义才能发展中国，
243.我们党和国家的中心任务是发展生产，
244.我们必须确立“科技兴国”的经济发展战略，
245.只要有发展前途，
246.特别是中日两国应该积极发展关系，
247.民族地区面临着极好的发展机遇，
248.“李鹏总理的报告为未来十年新疆的发展繁荣，
249.在稳定中发展中国，
250.这对于一个新的产业体系的建立和发展来说，
251.预测发展趋势，
252.大力调整了指导思想和发展战略，
253.山东的有关部门就全省的经济发展速度，
254.山东在今后十年已进一步明确了以效益为中心的发展战略，
255.联合发展经济，
256.深圳市已设立经济合作发展基金，
257.今年的“两会”描绘今后十年的发展蓝图，