Incremental Chinese Lexicon Extraction with Minimal Resources on a Domain-Specific Corpus

Gaël Patin
(1) Texts, Computer Science and Multilingualism Research Center (Ertim)
National Institute of Oriental Languages and Civilizations (Inalco)
(2) Arisem, Thales Company
gael.patin@arisem.com

Abstract

This article presents an original lexical unit extraction system for Chinese. The method is based on an incremental process driven by an association score featuring a minimal resources statistically aided linguistic approach. We also introduce a linguistics-based lexical unit definition and use it to describe an evaluation protocol dedicated to the task. The experimental results on a domain specific corpus show that the method performs better than other approaches. The extraction results, evaluated on a random sample of the working corpus, show a recall of 68.4 % and precision of 37.1 %.

1 Introduction

Lexical resources are all the more fundamental to NLP systems since domain specific corpora are multiple and various. The performance of common tasks, such as Information Retrieval or Information Extraction, can be improved by comprehensive and updated domain specific lexicon (i.e. terminology). However the constitution of lexicons raises pragmatic issues, such as development cost or re-usability, which have a great importance in an industrial context ; and also theoretical issues, such as the definition of the lexical unit or evaluation protocol, which are crucial for the relevance of the results. In Chinese text processing context, lexicons are particularly important for dictionary-based word segmentation techniques in which out-of-vocabulary words are an important cause of errors (Sproat and Emerson, 2003).

In this paper we consider the lexicon extraction task independent of the word segmentation, this position differs from Zhao and Kit's (2004) point of view. Generally speaking, word segmentation aims at delimiting units in a sequence of characters. The delimited units are usually morphological lexical units (i.e. words) and internal composition of the unit is not considered. The evaluation process checks whether each word occurrence is well delimited. On the opposite, lexicon extraction aims at extracting lexicon entries from a corpus. The extracted units are morphological or syntactic units and the internal components are also considered. The evaluation process checks the extracted candidates list considering the corpus global scope.

Many approaches for Chinese lexicon extraction rely on a supervised word segmenter (Wu and Jiang, 2003; Li et al., 2004) or a morpho-syntactic tagger (Piao et al., 2006) to extract unknown words. These techniques perform well but suffer from a major drawback, they cannot be applied efficiently to corpora that cover different domains than the calibration corpus. Some approaches are nested in an unsupervised word segmentation process and aim at improving its effectiveness. Fung and Wu (1994) try to select segments using mutual information on bigram. Chang and Su (1997) present an iterative unsupervised lexicon extraction system driven by the quality of segmentation obtained with the discovered lexicon. This approach, although efficient, imposes an arbitrarily 4-character length restriction on candidates. Other works, like this approach, focus on the lexicon or terminology extraction as standalone task. Feng et al. (2004) introduce a lexicon extraction unsuper-

Coling 2008: Poster Volume, pages 963–971,
Beijing, August 2010

vised method based on context variation with very convincing results. Yang et al. (2008) focus on terminology extraction using delimiters extracted from a training corpus with good results.

This study proposes an original answer to the Chinese lexicon extraction task using an incremental minimal resources method to extract and rank lexical unit candidates. An annotated reference corpus is required to extract a common-word dictionary and to prepare the data. The method has the advantage of proposing structured candidates, which allow interactive candidate filtering. In addition the candidate maximum length is determined by the number of associations that allow the detection of the longer lexical units. We extend the association measure method introduced by Sun et al. (1998) for word segmentation without lexical resources. This paper starts with a linguistic definition of the lexical unit which drives the method. We also build on it to propose an improvement of the evaluation protocol for the Chinese lexicon extraction task.

2 Lexical Unit Definition

Although defining the Chinese lexical unit is not a trivial task, we think that it is absolutely necessary for the understanding of the kind of linguistic phenomena we are dealing with. Without this knowledge we may miss important features and may not be able to efficiently evaluate the extraction process. We introduce two linguistic concepts to define the lexical units focusing on contemporary written Chinese: the *morpho-syntactic unit* and the *lexical content*. These definitions use concepts introduced by Polguère (2003) applied to the Chinese case by Nguyen (2008).

2.1 Morpho-syntactic Unit

A *graphy* is the Chinese minimal autonomous orthographic unit and it approximatively matches the glyph concept in computer science. The following glyphs are different Chinese graphies: 猫, 貓, 寿, 葡, 萄. A *morph* (noted | m |) is the smallest meaningful unit representable by a sequence of graphies. Morphs are atomic so that they cannot be representable by a smaller sequence of morphs. The following sequences of graphies are different morphs : |longevity 寿|, |grape 葡萄|, |aspirin 阿司匹林|, |buy 买|. Note that the graphy 萄 does not carry any meaning and is not a morph. A *morpheme* (noted ‖M‖[1]) is a set of morphs sharing the same lexical content ignoring grammatical inflection or variants (Table 1). Chinese morphs cannot be inflected, unlike European languages, but some graphies have variants.

Morpheme	***Morph***
‖protect 保‖	|protect 保|
‖aspirin 阿司匹林‖	|aspirin 阿司匹林|
‖cat 猫‖	|cat 猫| |cat 貓|

Table 1: *Morphemes and related morphs*

A *word-form* (noted (w)) is an autonomous and inseparable sequence of morphs. Autonomy means that it can be enunciated individually and can take place in a syntactic paradigm. Inseparability means that breaking the sequence causes the loss of the relationship between elements. A *lexeme* (noted ⟪ w ⟫) is a set of word-forms sharing the same lexical content ignoring inflection or variants (Table 2).

Lexeme	***Word-form***
⟪aspirin 阿司匹林⟫	(|aspirin 阿司匹林|)
⟪take 拿⟫	(|take 拿|) (|take /prefect/ 拿|了|) (|take /progressive/ 拿|着|) (|take /experience/ 拿|过|)
⟪insurance 保险⟫	(|insurance 保|险|)
⟪|panda 熊|猫|⟫	(|panda 熊|猫|) (|panda 熊|貓|)

Table 2: *Lexemes and associated word-forms*

A *phrase* (noted [s]) is a syntactic combination of word-forms. The syntactic nature of the combination implies that the phrase components are relatively free. A *locution* (noted ⟦ s ⟧) is a set of lexicalized phrases sharing the same lexical content ignoring inflection or variants (Table 3).

Locution	***Phrase***
⟦shoot 开枪⟧	[(shoot 开)(枪)] [(shoot /prefect/ 开了)(枪)] ...
⟦be jealous 吃醋⟧	[(be jealous 吃)(醋)]
⟦insurance company 保险公司⟧	[(insurance company 保险)(公司)]

Table 3: *Locutions and associated phrases*

[1]The standard simplified form is used to represent morphemes.

The morphs, word-forms and phrases are the morpho-syntactic units, they describe the composition of lexemes and locutions.

2.2 Lexical Content

The lexical units we look for are lexemes and locutions. Finding lexical units means identifying words-forms and phrases having a lexical content. We use two criteria to define the lexical content: the *compositionality criterion* and the *referentiality criterion* (Table 4). Units which fulfill at least one of these criteria are said to have a lexical content.

The compositionality criterion (or lexicalization criterion) is relative to the relationship between the sense of the unit and the sense of its components. The question is whether or not the sense of the unit can be deduced from the combination of its components. The referentiality criterion is related to the relationship between the unit and the referent concept or object. The question is whether or not the referent has specific properties for the speakers. This criterion is strongly dependent on human judgment and the working domain.

	Referential	***No-Referential***
Compositional	《Chinese food 中餐》	(anticization 古代化)
	⟦insurance company 保险公司⟧	[African car 非洲汽车]
Lexicalized	《disinfect 消毒》	⟦everyone 大家⟧
	⟦dividend product 分红产品⟧	⟦selling vinegar as wine 挂羊头卖狗肉⟧

Table 4: *Referential and Compositional units*

The Table 4 presents examples of four criterion combinations. Referentiality and compositionality criteria are always applied at the highest association level, thus ⟦insurance company 保险公司⟧) is compositional, although 《insurance 保险》 and 《company 公司》 are not compositional. Word-forms are not necessarily compositional or referential, thus the unit (anticization 古代化) does not refer to any concept and we can use the combination of its components to interpret it: 《antiquity 古代》 + ‖-ation 化‖. Referentiality does not imply lexicalization, thus the compositional unit ⟦German car 德国汽车⟧ is referential because it refers to the German car brands or characteristics in the automobile context.

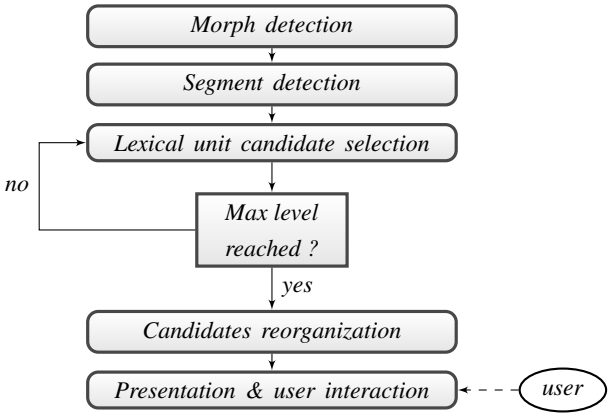

Figure 1: *Method overview*

3 Methodology

The method (Figure 1) follows the linguistic intuitions developed in the previous section. We identify morpho-syntactic units and select those that are likely to have a lexical content to obtain *lexical unit candidates* (LUCs). The word-forms and phrases are respectively generated by associations of morphs or word-forms and association of word-forms or phrases. We consequently use an incremental process, which associates LUCs as they are selected. The incremental process is initiated by detecting every morph and spliting the corpus into segments. Then we enumerate all the morpho-syntactic unit couples and use lexical content criteria to select the couples to associate. This process is repeated until the maximum number of associations is reached. At the end, the LUCs are reorganized and submitted to the user. The user's answers are used to filter the remaining LUCs.

3.1 Morphs Detection

As stated in Section 2.1, we consider that the morph is the minimal morpho-syntactic unit. Every glyph is considered as a morph unless it can be included in an ancient loanword morph (《butterfly 蝴蝶》 《garbage 垃圾》) or a foreign transcription morph (《Italy 意大利》, 《microphone 麦克风》). In an ambiguous case the longest possibility is accepted. Foreign transcriptions are phonetic interpretations of foreign words using the pronunciation of the Chinese graphies. The set of graphies used for transcription is well-

known and closed. We trained a CRF tagger[2] using simple features based on current, next and previous graphies to extract foreign transcriptions (the training corpus is described in Section 4.1). Ancient loanwords importation process is not productive anymore, thus they are detected using a loanword list.

3.2 Segment Detection

The aim of the segment detection step is to split the corpus into segments (i.e. a succession of Chinese graphies). Chinese texts contain two kinds of delimiters which are not likely to be components of a lexical unit, delimiter-words and delimiter-expressions. Delimiter-words are enumerable with a common word dictionary[3] and include prepositions (对, 使), adverbs (很, 也, 都), pronouns (我, 其他, 那儿), interrogative pronouns (哪里, 谁), conjunctions (而且, 但, 因此), discourse structure words (目前, 按照, 由), tonal particles (啊, 吧) and tool-words (的). Delimiter-expressions include numerical expressions (六万美元, 三个), temporal expressions (今天晚上, 八点左右), circumstantial expressions (从...开始, 在...中) , which are easily describable using shallow context-free grammars. Delimiters are removed from the corpus and used to delimit the segments. The inflexions (了, 过, 着), which introduce inflectional variations, are also removed from the corpus but do not delimit the segments. The delimiters identification is controlled by rules. For instance tonal particles are removed only if they are the end of a segment, discourse structure words are removed only if they are the beginning of a segment. Delimiters and inflexions are not removed if they are inside a sequence of graphies which is present in a common-word dictionary.

3.3 Selection of Lexical Unit Candidates

In this step, *lexical unit candidates* (LUCs) are extracted by selecting morpho-syntactic unit couples, which are likely to have a lexical content. The first assumption is that lexical units can always be decomposed into binary trees. Only a small number of lexical units do not satisfy this assumption (e.g. 乌漆墨黑), in such case it is possible to select a non-linguistically motivated way to decompose the unit into binary associations. Thus, every couples of contiguous morpho-syntactic units are iteratively enumerated for each segment. The second assumption is that *association measures* are good statistical evidence to detect lexical content. Thereby, the association strength of morpho-syntactic couples is used as a main criterion to identify relevant candidates.

Sentence with delimiters noted {delimiter}:

公司银代主力产品“新红A”、“新红C”两款分红产品适应{了}今年资本市场{的}现状，产品设计、分红水平、特殊红利分配{等}方面{都}得到合作银行{和}客户{的}认同，充分满足{了}客户{的}预期利益，{在}市场{上}得到{了}{很}高{的}美誉度。

Obtained segments noted [segment]:

[公司银代主力产品] [新红] [新红] [两款分红产品适应今年资本市场] [现状] [产品设计] [分红水平] [特殊红利分配] [方面] [得到合作银行] [客户] [认同] [充分满足客户] [预期利益] [市场] [得到] [高] [美誉度]

Figure 2: *Segment detection example*

Consider G the alphabet of all Chinese graphies, $M = G^+$ the language describing the morpho-syntactic units, S_n a set of segments at step n, $s_n^i = m_1, m_2, ..., m_n$ the i^{th} segment of S_n where $\forall m \in s_n^i \mid m \in M$ and S_n^* the set of all morpho-syntactic unit couples in S_n segments. Given the morpho-syntactic unit couple m_i, $m_{i+1} \in S_n^*$ (denoted as $m_{i,i+1}$), the *lexical content criteria* ($LCC(m_{i,i+1})$) matches if the following conditions are fulfilled:

1. Neither m_i nor m_{i+1} has not been associated at the current step n.
2. $Nb(m_i) \neq 1$ or $Nb(m_{i+1}) \neq 1$.
3. $AS(m_{i,i+1}) > T$.
4. $AS(m_{i,i+1}) > AS(m_{i-1,i})$ or not $LCC(m_{i-1,i})$
5. $AS(m_{i,i+1}) > AS(m_{i+1,i+2})$ or not $LCC(m_{i+1,i+2})$

where $Nb(x)$ is the number of occurrences of x, $AS(x, y)$ returns the association score of the cou-

[2] CRF++ implementation of Conditional Random Fields

[3] We assert that this kind of dictionary is easily available

ple x, y computed with a given association measure, and T is the association threshold relative to the association measure (cf. 4.1).

Let S_0 the initial set of segments where $\forall s_0^i \in S_0$, s_0^i is a segment (cf. 3.2) such that $\forall m \in s_0^i$, m is a morph (cf. 3.1). The LUC list is composed of morpho-syntactic couples produced by the association operator $\oplus$ to compute S_{max} (algorithm 1) with max the maximum number of iteration.

$$
\begin{aligned}
&S \leftarrow S_{n-1} \\
&\textbf{while} \quad \exists\, m_{i,i+1} \in S^* |\ LCC(m_{i,i+1}) \\
&\quad S \leftarrow S[m_i \oplus m_{i+1}] \qquad (1) \\
&\textbf{end} \\
&S_n \leftarrow S
\end{aligned}
$$

with $\oplus$ the association operator whose result is a morpho-syntactic unit, $S_n[m_1 \oplus m_2]$ the replacement of m_1 and m_2 by the morpho-syntactic unit $m_1 \oplus m_2$ in the corresponding segment. See the Section 5 for more details about the maximum number of iteration setting.

3.4 Candidates Reorganization

Once LUCs are extracted, we map every LUC to the couple of morpho-syntactic units it is composed of. These units are called *components*. Some LUCs are generated from two different couples at the candidate selection step. For instance, 旅游业者 is discovered in two ways: 旅游⊕业者 or 旅游业⊕者. We always choose the most frequent option. When the "LUC/couple" map is created, we sort the LUCs by their corresponding couple association scores. Finally, if a LUC is ranked in the list before its components we move the components to the position just before it in the list and use the same rule to recursively check the moved components. The candidates list is expected to be ordered by likelihood deduced by an association measure and compositional order.

3.5 Presentation and User Interaction

The lexicon extraction task aims at submitting a ranked list of candidates to the user in order to help him produce lexical resources. The user is expected to check the list in this order and the method uses the user answers to discard not yet verified candidates. To do so, the user is asked to answer the following questions for each LUC according to the definition given in the Section 2:

1. Does the unit have a lexical content ?
2. Is the unit a part of a lexical unit ?

If answers to both these questions are 'no' then all the candidates having this component are removed from the remaining list.

4 Evaluation

Since the submitted candidates are progressively modified according to the user answers, the evaluated candidates are only the ones submitted to the user. We used three measures to evaluate the method: recall, precision and precision at rank n. Since producing large annotated corpora is costly, we perform the evaluation using a sample of texts from the evaluation corpus. Therefore the scores obtained are an estimation of the true scores. The inter-human variation is not considered here and should be integrated in further works.

4.1 Evaluation parameters

The morphs and the segment detection step use data from a *reference corpus*: *The Lancaster Corpus of Mandarin Chinese* (McEnery and Xiao, 2004). The corpus is composed of text samples choose in various domain and genre corpora, it contains two millions of glyphs and it is annotated according to the Beijing University annotation guideline[4]. This corpus is mainly used to extract delimiter-words, to produce the grammar for delimiter-expressions and to extract a common-word dictionary. All foreign transcriptions are also annotated for the CRF tagger training (cf. 3.1).

The lexical unit detection step is evaluated using four well-known association measures: Pointwise Mutual Information (PMI), Poisson-Striling (PS) (Quasthoff and Wolff, 2002), Log-likelihood (LL), Pointwise Mutual Information Cube (PMI^3) (Daille, 1994). These measures are detailed in table 5. The significant association threshold is intuitively given by the statistical interpretation of

[4] *http://icl.pku.edu.cn/icl_groups/corpus/coprus-annotation.htm*

AM	*Formulas*	*Variables*
PMI	$\log \frac{p_{xy}}{p_{x*}p_{*y}}$	x, y : words $\overline{x}$: all words but x $*$: all words p_x : x probability f_x : x frequency N : nb. of bigram $\lambda = N \cdot p_x \cdot p_y$ $k = f_{xy}$ $\hat{f}_{xy} = \frac{f_{x*}f_{*y}}{N}$
LL	$2 \sum_{i,j}^{\{x,\overline{x}\},\{y,\overline{y}\}} f_{ij} \log \frac{f_{ij}}{\hat{f}_{ij}}$	
PS	$\frac{k(\log k - \log \lambda - 1)}{\log N}$	
PMI^3	$\log \frac{N {f_{xy}}^3}{f_{x*}f_{*y}}$	

Table 5: *Association score calculation*

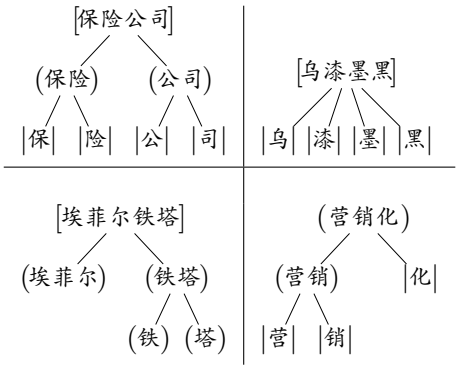

Figure 3: *Lexical unit trees*

the formulas for MI and PS. Thus, these measures are used for LCC's *selection criterion 2* and T is set to 0 (cf. 3.3). A threshold can not be deduced from PS and PMI^3, therefore they are only used for LCC's *comparison criteria 3 & 4*.

4.2 Evaluation Process

To prepare the evaluation we randomly selected twenty texts in an evaluation corpus and annotated lexical units according to the linguistic description given in Section 2. For each sample text, we obtained a set of lexical unit trees (Table 3) corresponding to all the encountered lexical units. N-trees are used for units which can not be transformed into binary tree. Two evaluation sets are defined, the *shallow set* which contains the root nodes of the lexical unit trees and the *deep set* which contains the inner nodes[5] of the lexical unit trees. Given the four examples of Figure 3, the shallow set contains [保险公司], [乌漆墨黑], [埃菲尔铁塔] and (营销化) ; and the deep set contains [保险公司], (保险), (公司), [乌漆墨黑], [埃菲尔铁塔], (铁塔), (营销化) and (营销).

Experiments with different parameters produce different candidate lists and an expert intervention is required to evaluate each candidate list. To avoid this problem, all the repeated sequences of non-inflectional graphies are generated from the annotated sample texts and intersected with the LUC list. The obtained list is a projection of the candidate list on the sample texts. This trick allows us to extract all LUCs appearing in the sample texts and evaluate them automatically.

[5] All nodes excluding leaves.

5 Experiments

The experiments are conducted on insurance domain corpus containing ten million graphies. This evaluation corpus is composed of news and articles collected automatically from Chinese insurance companies websites. The text fields are extracted with an xhtml parser. Several text fields, such as menus or buttons, are repeated and duplicates are removed to avoid noise. The presented method, referred as ILex (Incremental Lexicon Extractor), is applied using the previously mentioned 4 measures (cf. 4.1). The evaluation is based on couple of measures, the first measure is dedicated to candidates selection (LCC 2.) and the second to candidates comparison (LCC 3. & 4.). The comparison measure is also used to sort the candidates (cf. 3.4). The maximal number of iterations is set to 3 (for a maximal depth of 4), which is the maximum number of associations required to compose the majority of lexical units in the reference corpus. The precision and recall are computed on the deep set in order to consider all valid lexical units, the recall on the shallow set is given to see the results on wider lexical units (Table 6). The results show that PMI-LL couple performs better overall than the other measures. It can be noticed that the scores are relatively close ($\pm$1.8% for precision and $\pm$7.0% for deep recall) meaning that the choice of the association measure has a low influence over the results. For the further experiments are conducted with PMI-LL, which achieves the best recall score.

Selection	PMI		PS	
Comparison	LL	PMI^3	LL	PMI^3
Precision	37.1	38.9	37.3	38.1
Deep recall	68.4	65.6	62.3	61.4
Shallow recall	75.1	74.2	70.6	70.6

Table 6: *Measure combinations results*

The method extracted 585,794 LUCs from the whole corpus using the PMI-LL couple before applying the user interaction step. The *candidate list projection* (cf. 4.2) contains 4,539 LUCs. The user decisions are simulated with the lexical unit trees obtained from sample texts. In total 312 LUCs were removed in consequence of the user interaction (cf. 3.5), without this step the precision decreases to 33.7%. The 1,246 LUCs present in the common-word dictionary are ignored. Finally 1,886 invalid candidates and 1,105 valid lexical units are submitted to the user, the evaluation is based on these 3,059 LUCs.

Lexical unit	**Rank**	**Nb.**
⟦政策性农业保险⟧ (policy agricultural insurance)	155	1798
⟦泰康人寿⟧ (Tai Kang Life Insurance)	453	1,854
⟪保险人⟫ (insurer)	1,048	4,999
⟦南开大学⟧ (Nan Kai University)	2,828	111
⟦洛杉矶旅游业者⟧ (Los Angeles tourism professionals)	9,647	3
⟦人寿保险⟧ (life insurance)	11,647	871
⟪王恩韶⟫ (Wang Enshao (person))	14,617	2
⟦有偿使用⟧ (compensated use)	34,596	8
⟪太湖流域⟫ (Taihu Lake Basin)	102,612	2
⟪择机⟫ (wait an opportunity)	126,044	31
⟦中华人民共和国劳动合同法⟧ (The People's Republic of China labor contract law)	387,235	1

Table 7: *Sample of extracted lexical units*

A sample of extracted lexical units is presented in Table 7. In this list, the lower number of occurrences is 1 and the longest unit length is 12 graphies. Most of the extracted lexical unit are terms, a significant number of people names, common words and larger named entities are extracted too. The most part of the very frequent lexical units are ranked at the top of the list but some low frequency LUCs are ranked over the high frequency candidates. The Figure 4 presents

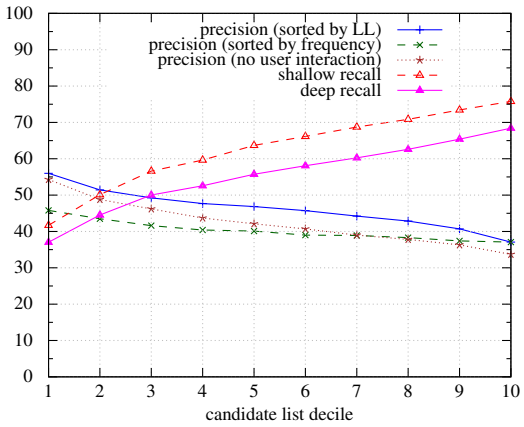

Figure 4: *ILex results using PMI-LL*

the results as a function of the LUC list deciles. The LL sorting is compared to frequency sorting for the precision at rank n. The LL sorting curve is above the frequency sorting curve, this fact shows that LL is more efficient at sorting valid LUCs. The majority of the missed candidates have a low number of occurrences ($\leq$3) and 57.8% of the longest lexical unit ($>$7) are also missed. Most of extraction errors have a low number of occurrences, 40.1% of the errors are caused by lexical unit composition errors (e.g. $*(\text{保险})\oplus|\overset{\text{study}}{\text{学}}|$ in $[(\text{保险})(\overset{\text{insurance institute}}{\text{学院}})]$ or $*(\overset{\text{reform}}{\text{改革}})|\overset{\text{commission}}{\text{委}}|$ in $[(\overset{\text{reform \& development commission}}{\text{发展}})(\text{改革})|\text{委}|]$) and 59.9% by association errors (e.g. $(\overset{\text{extend}}{\text{扩大}})\oplus[\overset{\text{agricultural Insurance}}{\text{农业保险}}]$ or $(\overset{\text{standard}}{\text{规范}})\oplus(\overset{\text{development}}{\text{发展}})$).

The AccessVar method (Feng et al., 2004), an unsupervised lexicon extraction method having the best performance, was reimplemented and used as a reference. This method uses the corpus substrings' number of distinct contexts, noted AV (*accessor variety*), to extract candidates. AccessVar is configured by an *accessors variety threshold* (AVT), which is the minimal AV required to hold a candidate, the number of occurrences of candidates is consequently greater or equal to the AVT. For the experiments, the candidate maximal length is set to 7 graphies[6] and AVT to 3. Similarly, ILex candidates appearing less than three times and having a length greater than 7 are discarded. The ILex user interaction is not applied for this comparison. In order unify the input data, AccessVar handles the segments detected by ILex

[6]Higher values cause space complexity issues.

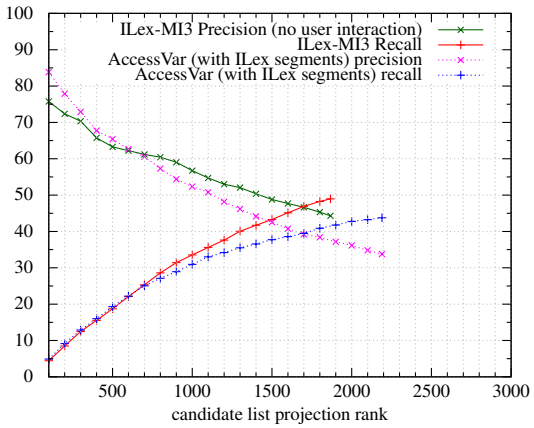

Figure 5: *ILex & AccessVar results*

instead of the corpus full text.

AccessVar and ILex extract respectively 125,467 and 116,412 LUCs and the candidate list projection contains 2,190 and 1,876 LUCs. The results are computed on the deep set (figure 5). AccessVar and ILex achieve respectively recall of 43.7% and 49.0%. A total of 667 of the lexical units extracted are common to both methods, 161 lexical units are extracted exclusively by ILex and 74 lexical units are extracted exclusively by AccessVar. This means that both methods have close covering capacities. From rank 100 to rank 700, the results are close but the curves begin to diverge after this rank, this trend means that the performance are similar for the 700 best candidates. However, ILex achieves 44.4% precision which is 10.6% higher than AccessVar (33.8%), this difference, in view of the close recall score, shows that ILex generates less invalid candidates. The errors specific to AccessVar are due to context adhesion errors (e.g. $*(\overset{\text{company}}{\text{保险}})\oplus|\overset{\text{specialty}}{\text{产}}|$ in $[(\overset{\text{insurance industry}}{\text{保险}})(\text{产业})]$, $[(\overset{\text{insurance product}}{\text{保险}})(\text{产品})]$, $[(\overset{\text{insurance produce}}{\text{保险}})(\text{产生})]$ etc.), or association errors (e.g. $*|\overset{\text{country}}{\text{国}}|\oplus|\overset{\text{East}}{\text{东}}|$, $*(\overset{\text{industry}}{\text{工业}})\oplus(\overset{\text{group}}{\text{集团}})$). ILex avoids these errors because of three mechanisms. First, the statistical likelihood between the couple components is tested (e.g. $*|\overset{\text{country}}{\text{国}}|\oplus|\overset{\text{East}}{\text{东}}|$ PMI score is negative). Second, the method checks association likelihood of the contexts before associating two morpho-syntactic units, (e.g. $(\overset{\text{aeronautic}}{\text{航空}})(\overset{\text{industry}}{\text{工业}})$ score is over $*(\overset{\text{industry}}{\text{工业}})\oplus(\overset{\text{group}}{\text{集团}})$ score in $[\overset{\text{Aviation Industries Corporation of China}}{\text{中国航空工业集团公司}}]$). Third, the incremental association process determine smaller unit before trying associating bigger couples (e.g. $(\overset{\text{insurance}}{\text{保险}})$ and $(\overset{\text{industry}}{\text{产业}})$ are associated before $[\overset{\text{insurance industry}}{\text{保险产业}}]$).

6 Conclusion and Further Works

The presented method features incremental lexical unit extraction with interactive candidate filtering capability. The maximal candidate length is not imposed directly, but instead is determined by the maximal number of associations. The lexical resources required are re-usable and non-domain specific, which significantly reduce their cost for long-term deployment. The method achieves decent performance and improves the reference method's precision for this task. Furthermore, the extracted results include low-frequency and long candidates which are known to be difficult to extract. Finally, the binary association process allows us to sort the candidates by association measure, which is more relevant than frequency.

This paper also introduced the beginning of a linguistically consistent lexical unit definition. This definition draws the outlines of a corpus annotation guide dedicated to the lexicon extraction task. The evaluation process is improved by the lexical unit trees annotations and a candidate list projection technique, which allows full-automatic estimation of extraction system performance.

The first upcoming objective is the development of a robust evaluation protocol for the lexical extraction task. This is crucial for further improvements and means that the variation between annotators of the evaluation corpus, and the stability of the method over different corpora need to be considered. Finally we will try to solve the not yet managed lexicon extraction issues, Latin characters tokens which cause the method miss some extractions (e.g. $(\!(\overset{\{\textit{product name}\}}{\text{新红A}})\!)$), and the discontinuous locutions (e.g. $[\![\overset{\text{reach on the phone}}{\text{打通电话}}]\!]$ in $(\overset{\text{call / through}}{\text{打通了}})\text{常总的}(\overset{\text{phone}}{\text{电话}})$ or $[\![\overset{\text{bear responsibility}}{\text{负责任}}]\!]$ in $\text{本公司}(\overset{\text{bear}}{\text{负}})\text{给付保险金}(\overset{\text{responsibility}}{\text{责任}})$).

Acknowledgements

Our sincere thanks to the anonymous reviewers. Special thanks to Pierre Zweigenbaum, to all my colleagues from Arisem and Ertim and to the corpus annotators without which this work would not be possible.

References

Chang, Jing-Shin and Keh-Yih Su. 1997. An unsupervised iterative method for Chinese new lexicon extraction. *International Journal of Computational Linguistics & Chinese Language Processing*, vol. 1(1), pp. 101–157.

Daille, Béatrice. 1994. *Approche mixte pour l'extraction automatique de terminologie: statistiques lexicales et filtres linguistiques*. PhD thesis, Université Paris 7.

Feng, Haodi, Kang Chen, Xiaotie Deng and Weimin Zheng. 2004. Accessor variety criteria for Chinese word extraction. *Computational Linguistics*, vol. 30:1, pp. 75-93.

Fung, Pascale and Dekai Wu. 1994. Statistical augmentation of a Chinese machine-readable dictionary. In *WVLC-2, Second Annual Workshop on Very Large Corpora (COLING-94)*, Kyoto, Japan, pp. 69-85.

Hai, Zhao and Chunyu Kit. 2008. An Empirical Comparison of Goodness Measures for Unsupervised Chinese Word Segmentation with a Unified Framework. *In Proceedings of the Third International Joint Conference on Natural Language Processing (IJCNLP-08)*, Hyderabad, India, Vol. 1, pp. 9-16.

McEnery, Tony and Richard Xiao. 2004. The Lancaster Corpus of Mandarin Chinese: A corpus for monolingual and contrastive language study. *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC-04)*, Lisbon, Portugal, pp. 1175-1178.

Li, Hongqiao, Changning Huang, Jiangfen Gao and Xiaozhong Fan. 2004. The use of SVM for Chinese new word identification. *First International Joint Conference on Natural Language Processing (IJCNLP-04)*, Sanya, China, pp. 497-504.

Nguyen, Etienne Van Tien. 2008. *Unité lexicale et morphologie en chinois mandarin – vers l'élaboration d'un DEC du chinois*. Phd thesis, Montréal University.

Piao, Scott S. L., Guangfan Sun, Paul Rayson and Qi Yuan. 2006. Automatic extraction of Chinese multiword expressions with a statistical tool. *Workshop on Multi-word-expressions in a Multilingual Context held in conjunction with the 11th EACL*, Trento, Italy, pp. 17-24.

Polguère, Alain. 2003. *Lexicologie et sémantique lexicale. Notions fondamentales*. Presses de l'Université de Montréal, Coll. Paramètres.

Quasthoff, Uwe and Christian Wolff. 2003. *The Poisson collocation measure and its application*. In *Second International Workshop on Computational Approaches to Collocations*, Vienna, Austria.

Sproat, Richard and Tom Emerson. 2003. The first international Chinese word segmentation bakeoff. *Proceedings of the Second SIGHAN Workshop on Chinese Language Processing*, Japan, vol. 17, pp. 133-143.

Sun, Maosong, Danyang Shen and Benjamin K Tsou. 1998. Chinese Word segmentation without lexicon and hand-crafted training data. *Proceedings of the 17th international conference on Computational linguistics*, Montreal, Canada, Vol. 2, pp. 1265-1271.

Wu, Andi and Zixin Jiang. 2000. Statistically-enhanced new word identification in a rule-based Chinese system. *Proceedings of the 2nd Chinese Language Processing Workshop*, Hong-Kong, vol. 12, pp. 45-51.

Yang, Yuhang, Qin Lu and Tiejun Zhao. 2008. Chinese Term Extraction Using Minimal Resources. *Proceedings of the 22nd International Conference on Computational Linguistics*, Manchester, United Kingdom, Vol. 1, pp.1033-1040.