A Bio-inspired Approach for Multi-Word Expression Extraction

Jianyong Duan, Ruzhan Lu
Weilin Wu, Yi Hu
Department of Computer Science
Shanghai Jiao Tong University
Shanghai, 200240, P.R. China
duanjy@hotmail.com
{lu-rz,wl-wu,huyi}@cs.sjtu.edu.cn

Yan Tian
School of Foreign Languages
Department of Computer Science
Shanghai Jiao Tong University
Shanghai, 200240, P.R. China
tianyan@sjtu.edu.cn

Abstract

This paper proposes a new approach for Multi-word Expression (MWE)extraction on the motivation of gene sequence alignment because textual sequence is similar to gene sequence in pattern analysis. Theory of Longest Common Subsequence (LCS) originates from computer science and has been established as affine gap model in Bioinformatics. We perform this developed LCS technique combined with linguistic criteria in MWE extraction. In comparison with traditional n-gram method, which is the major technique for MWE extraction, LCS approach is applied with great efficiency and performance guarantee. Experimental results show that LCS-based approach achieves better results than n-gram.

1 Introduction

Language is under continuous development. People enlarge vocabulary and let words carry more meanings. Meanwhile the language also develops larger lexical units to carry specific meanings; specifically MWE's, which include compounds, phrases, technical terms, idioms and collocations, etc. The MWE has relatively fixed pattern because every MWE denotes a whole concept. In computational view, the MWE repeats itself constantly in corpus(Taneli,2003).

The extraction of MWE plays an important role in several areas, such as machine translation (Pascale,1997), information extraction (Kalliopi,2000) etc. On the other hand, there is also a need for MWE extraction in a much more widespread scenario namely that of human translation and technical writing. Many efforts have been devoted to the study of MWE extraction (Beatrice,2003; Ivan,2002; Jordi,2001). These statistical methods detect MWE by frequency of candidate patterns. Linguistic information as a filtering strategy is also performed to improve precision by ranking their candidates (Violeta,2003; Stefan,2004; Arantza,2002). Some measures based on advance statistical methods are also used, such as mutual expectation with single statistic model (Paul,2005),C-value/NC-value method (Katerina,2000),etc.

Frequent information is the original data for further MWE extraction. Most approaches adopt n-gram technique(Daniel,1977; Satanjeev,2003; Makoto,1994). n-gram concerns about one sequence for each time. Every sequence can be cut into some segments with varied lengths because any length of segment has the possibility to become candidate MWE. The larger the context window is, the more difficulty its parameters acquire. Thus data sparseness problem deteriorates. Another problem arises from the flexible MWE which can be separated by an arbitrary number of blanks, for instance, "make. . . . . decision". These models cannot effectively distinguish all kinds of variations in flexible MWE.

On the consideration of relations between textual sequence and gene sequence, we propose a new bio-inspired approach for MWE identification. Both statistical and linguistic information are incorporated into this model.

2 Multi-word Expression

Multi-word Expression(in general, term) as the linguistic representation of concepts, also has some special statistical features. The component words of terms co-occur in the same context fre-

Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions, pages 176–182,
Sydney, July 2006.

quently. MWE extraction can be viewed as a problem of pattern extraction. It has two major phases. The first phase is to search the candidate MWEs by their frequent occurrence in the corpus. The second phase is to filter true MWEs from noise candidates. Filtering process involves linguistic knowledge and some intelligent observations.

MWE can be classified into strict patterns and flexible patterns by structures of their component words(Joaquim,1999). For example, a textual sequence s $= w_1 w_2 \cdots w_i \cdots w_n$, may contain two kinds of patterns:

Strict pattern: $\mathrm{p}_i = w_i w_{i+1} w_{i+2}$

Flexible pattern: $\mathrm{p}_j = w_i \sqcup w_{i+2} \sqcup w_{i+4}, p_k = w_i \sqcup \sqcup w_{i+3} w_{i+4}$

where $\sqcup$ denotes the variational or active element in pattern. The flexible pattern extraction is always a bottleneck for MWE extraction for lack of good knowledge of global solution.

3 Algorithms for MWE Extraction

3.1 Pure Mathematical Method

Although sequence alignment algorithm has been well-developed in bioinformatics (Michael,2003), (Knut,2000), (Hans,1999), it was rarely reported in MWE extraction. In fact, it also applies to MWE extraction especially for complex structures.

Algorithm.1.

1. Input:tokenlized textual sequences Q $= \{s_1, s_2, \cdots, s_n\}$
2. Initionalization : $pool,\ \Omega = \{\Omega_k\}, \Psi$
3. Computation:
   - I. Pairwise sequence alignment
     for all $s_i, s_j \in Q, s_i \neq s_j$
     $Similarity(s_i, s_j)$
     $Align(s_i, s_j) \overset{path(l_i,l_j)}{\longrightarrow} \{l_i,\ l_j,\ c_k\}$
     $pool \leftarrow\ pool \cup \{(l_i, c_k),\ (l_j,\ c_k)\}$
     $\Gamma \leftarrow \Gamma \cup c_k$
   - II. Creation of consistent set
     for all $c_k \in \Gamma,\ (l_i, c_k) \in pool$
     $\Omega_k \leftarrow \Omega_k + \{l_i\}$
     $pool \leftarrow\ pool - (l_i, c_k)$
   - III. Multiple sequence alignment
     for all Ω_k
     $star\_align(\Omega_k) \rightarrow MWU\ \Psi \leftarrow \Psi \cup MWU$
4. Output: Ψ

Our approach is directly inspired by gene sequence alignment as algorithm. 1. showed. The textual sequence should be preprocessed before input. For example, plurals recognition is a relatively simple task for computers which just need to check if the word accord with the general rule including rule (+s) and some alternative rules (-y + ies), etc. A set of tense forms, such as past, present and future forms, are also transformed into original forms. These tokenlized sequences will improve extraction quality.

Pairwise sequence alignment is a crucial step. Our algorithm uses local alignment for textual sequences. The similarity score between s$[1 \ldots i]$ and t$[1 \ldots i]$ can be computed by three arrays G$[i, j]$, E$[i, j]$,F$[i, j]$ and zero, where entry $\delta(x, y)$ means word x matches with word y; V$[i, j]$ denotes the best score of entry $\delta(x, y)$; G$[i, j]$ denotes s$[i]$ matched with $t[j]$:$\delta(s[i], t[j])$; $E[i, j]$ denotes a blank of string s matched with t$[j]$: $\delta(\sqcup, t[j])$; $F[i, j]$ denotes s$[i]$ matched with a blank of string t : $\delta(s[i], \sqcup)$.

Initialization:

$V[0, 0] = 0;\ V[i, 0] = E[i, 0] = 0;\ 1 \leq i \leq m.\ V[0, j] = F[0, j] = 0;\ 1 \leq j \leq n.$

A dynamic programming solution:

$$V[i, j] = \max\{G[i, j], E[i, j], G[i, j], 0\};$$

$$G[i, j] = \delta(i, j) + \max \begin{cases} G[i-1, j-1] \\ E[i-1, j-1] \\ F[i-1, j-1] \\ 0 \end{cases}$$

$$E[i, j] = \max \begin{cases} -(h+g) + G[i, j-1] \\ -g + E[i, j-1] \\ -(h+g) + F[i, j-1] \\ 0 \end{cases}$$

$$F[i, j] = \max \begin{cases} -(h+g) + G[i-1, j] \\ -(h+g) + E[i-1, j] \\ -g + F[i-1, j] \\ 0 \end{cases}$$

Here we explain the meaning of these arrays:

I. $G[i, j]$ includes the entry $\delta(i, j)$, it denotes the sum score is the last row plus the maximal score between prefix $s[1 \ldots i-1]$ and t$[1 \ldots j-1]$.

II. Otherwise the related prefixes $s[1 \ldots i]$ and t$[1 \ldots j-1]$ are needed[1]. They are used to check the first blank or additional blank in order to give appropriate penalty.

a. For $G[i, j-1]$ and $F[i, j-1]$, they don't end with a blank in string s. The blank $s[i]$ is the first blank. Its score is $G[i, j-1]$ (or $F[i, j-1]$) minus $(h+g)$.

b. For $E[i, j-1]$,The blank is the additional blank which should be only subtracted g.

In the maximum entry, it records the best score of optimum local alignment. This entry can be viewed as the started point of alignment. Then we backtrack entries by checking arrays which are generated from dynamic programming algorithm. When the score decrease to zero, alignment extension terminates. Finally, the similarity and alignment results are easily acquired.

Lots of aligned segments are extracted from pairwise alignment. Those segments with common component words (c_k) will be collected into the same set. It is called as consistent set for further multiple sequence alignment. These consistent sets collect similar sequences with score greater than certain threshold.

We perform star-alignment in multiple sequence alignment. The center sequence in the consistent set which has the highest score in comparison with others, is picked out from this set. Then all the other sequences gather to the center sequence with the technique of "once a blank, always a blank". These aligned sequences form common regions with n-column or a column. Every column contains one or more words. Calculation of dot-matrices is a widespread tool for common region analysis. Dot-plot agreement is developed to identify common patterns and reliably aligned regions in a set of related sequences. If several plots calculate consistently in a sequence set, it displays the similarity among them. It increases credibility of extracted pattern in this consistent set. Finally MWE with detailed pattern emerges from this aligned sequence set.

[1]Analysis approaches for $F[i, j]$ and $E[i, j]$ are the same, here only $E[i, j]$ is given its detailed explanation.

3.2 Linguistic Knowledge Combination

3.2.1 Heuristic Knowledge

Original candidate set is noise. Many meaningless patterns are extracted from corpus. Some linguistic rules (Argamon,1999) are introduced into our model. It is observed that candidate pattern should contain content words. Some patterns are only organized by pure function words, such as the most frequent patterns "the to", "of the". These patterns should be moved out from the candidate set. Filter table with certain words is also performed. For example, some words, like "then", cannot occur in the beginning position of MWE. These filters will reduce the number of noise patterns in great extent.

3.2.2 Embedded Base Phrase detection

Short textual sequence is apt to produce fragments of MWE because local alignment ends pattern extension when similarity score reduces to zero. The matched component words increase similarity score while unmatched words decrease it. The similarity scores of candidates in textual sequences are lower for lack of matched component words. Without accumulation of higher similarity score, pattern extension terminates quickly. Pattern extension becomes especially sensitive to unmatched words. Some isolated fragments are generated in this circumstance. One solution is to give higher scores for matched component words. It strengthens pattern extension ability at the expense of introducing noise.

We propose Embedded base phrase(EBP) detection as algorithm.2. It improves pattern extraction by giving lower penalty for longer base phrase. EBP is the base phrase in a gap (Changning,2000). It does not contain other phrase recursively. Good quality of MWE should avoid irrelative unit in its gap. The penalty function discerns the true EBP and irrelative unit in a gap only by length information. Longer gap means more irrelative unit. It builds a rough penalty model for lack of semantic information. We improve this model by POS information. POS tagged textual sequence is convenient to grammatical analysis. True EBP[2] gives comparatively lower penalty.

Algorithm.2.

1. Input: LCS of s_l, s_k

[2]The performance of our EBP tagger is 95% accuracy for base noun phrase and 90% accuracy for general use.

2. Check breakpoint in LCS

   i. Anchor neighbored common words and denote gaps
   for all $w_s = w_p, w_t = w_q$
   if $w_s \in l_s, w_t \in l_t, l_s \neq l_t$
   denote g_{st}, g_{pq}

   ii. Detect EBP in gaps
   $g_{st} \xrightarrow{EBP} g'_{st}, g_{pq} \xrightarrow{EBP} g'_{pq}$

   iii. Compute new similariy matrix in gaps
   $similarity(g'_{st}, g'_{pq})$

3. Link broken segment
   if $path(g'_{st}, g'_{pq})$
   $l_{st} = l_s + l_t, l_{pq} = l_p + l_q$

For textual sequence: $w_1 w_2 \cdots w_n$, and its corresponding POS tagged sequence: $t_1 t_2 \cdots t_n$, we suppose $[w_i \cdots w_j]$ is a gap from w_i to w_j in sequence $\cdots \; w_{i-1} [w_i \cdots w_j] w_j \; \cdots$. The corresponding tag sequence is $[t_i \cdots t_j]$. We only focus on EBP analysis in a gap instead of global sequence. Context Free Grammar (CFG) is employed in EBP detection. CFG rules follow this form:

$$
\begin{aligned}
&(1) EBP \leftarrow adj. + noun \\
&(2) EBP \leftarrow noun + "of" + noun \\
&(3) EBP \leftarrow adv. + adj. \\
&(4) EBP \leftarrow art. + adj. + noun \\
&\cdots
\end{aligned}
$$

The sequences inside breakpoint of LCS are analyzed by EBP detection. True base phrase will be given lower penalty. When the gap penalty for breakpoint is lower than threshold, the broken segment reunites. Based on experience knowledge, when the length of a gap is less than four words, EBP detection using CFG can gain good results. Lower penalty for true EBP will help MWE to emerge from noise pattern easily.

4 Experiments

4.1 Resources

A large amount of free texts are collected in order to meet the need of MWE extraction. These texts are downloaded from internet with various aspects including art, entertainment, military, business, etc. Our corpus size is 200, 000 sentences. The average sentence length is 15 words in corpus.

In addition, result evaluation is a hard job. Its difficulty comes from two aspects. Firstly, MWE identification for test corpus is a kind of labor-intensive business. The judgment of MWEs requires great efforts of domain expert. It is hard and boring to make a standard test corpus for MWE identification use. It is a bottleneck for large scales use. Secondly it relates to human cognition in psychological world. It is proved by experience that various opinions cannot simply be judged true or false. As a compromise way, gold standard set can be established by some accepted resources, for example, WordNet, as an online lexical reference system, including many compounds and phrases. Some terms extracted from dictionaries are also employed in our experiments. There are nearly 70,000 MWEs in our list.

4.2 Results and Discussion

4.2.1 Close Test

We created a closed test set of 8,000 sentences. MWEs in corpus are extracted by manual work. Every measure in both n-gram and LCS approaches complies with the same threshold, for example threshold for frequency is five times.Two conclusions are drawn from Tab.1.

Firstly, LCS has higher recall than n-gram but lower precision on the contrary. In close test set, LCS recall is higher than n-gram. LCS unifies all the cases of flexible patterns by GAM. However n-gram only considers limited flexible patterns because of model limitation. LCS nearly includes all the n-gram results. Higher recall decreases its precision to a certain extent because some flexible patterns are noisier than strict patterns. Flexible patterns tend to be more irrelevant than strict patterns. The GAM just provides a wiser choice for all flexible patterns by its gap penalty function. N-gram gives up analysis on many flexible patterns without further ado. N-gram ensures its precision by taking risk of MWE loss .

Secondly, advanced evaluation criterion can place more MWEs in the front rank of candidate list. Evaluation metrics for extracted patterns play an important role in MWE extraction. Many criteria, which are reported with better performances, are tested. MWE identification is similar to IR task. These measures have their own advantages to move interested patterns forward in the candidate list. For example, Frequency data contains much noise. True mutual infor-

Table 1: Close Test for N-gram and LCS Approaches

Measure	N-Gram			LCS		
	Precision (%)	Recall (%)	F-Measure (%)	Precision (%)	Recall (%)	F-Measure (%)
Frequency	35.2	38.0	36.0	32.1	48.2	38.4
TMI	44.7	56.2	49.1	43.2	62.1	51.4
ME	51.6	52.6	51.2	44.7	65.2	52.0
Rankratio	62.1	61.5	61.1	57.0	83.1	68.5

mation (TMI) concerns mutual information with logarithm(Manning,1999). Mutual expectation (ME) takes into account the relative probability of each word compared to the phrase(Joaquim,1999). Rankratio performs the best on both n-gram and LCS approaches because it provides all the contexts which associated with each word in the corpus and ranks them(Paul,2005). With the help of advanced statistic measures, the scores of MWEs are high enough to be detected from noisy patterns.

4.2.2 Open Test

In open test, we just show the extracted MWE numbers in different given corpus sizes. Two phenomena are observed in Fig.1.

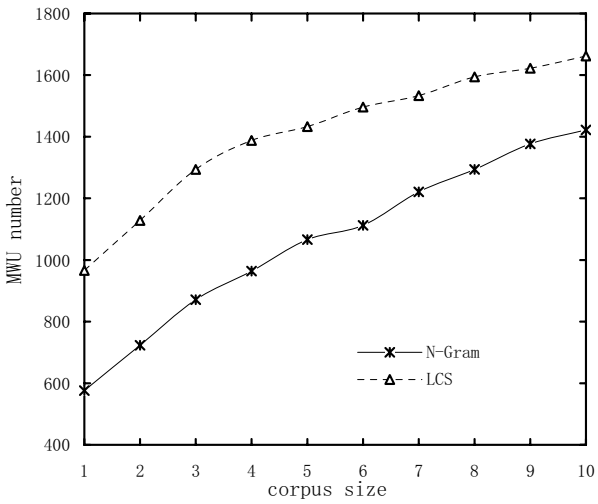

Figure 1: Open Test for N-gram and LCS Approaches

Firstly, with the enlargement of corpus size(every step of corpus size is 10,000 sentences), the detected MWE numbers increase in both approaches. When the corpus size reaches certain values, their increment speeds turn slower. It is reasonable on condition that MWE follow normal distribution. In the beginning, frequent MWEs are detected easily, and the number increases quickly. At a later phase, the detection goes into comparatively infrequent area. Mining these MWEs always need more corpus support. Lower increment speed appears.

Secondly, although LCS always keeps ahead in detecting MWE numbers, their gaps reduce with the increment of corpus size. LCS is sensitive to the MWE detection because of its alignment mechanism in which there is no difference between flexible pattern and strict pattern. In the beginning phase, LCS can detect MWEs which have high frequencies with flexible patterns. N-gram cannot effectively catch these flexible patterns. LCS detects a larger number of MWE than n-gram does. In the latter phase, many variable patterns for flexible MWE can also be observed, among which relatively strict patterns may appear in the larger corpus. They will be catched by n-gram. On the surface of observation, the discrepancy of detected numbers is gradually close to LCS. In nature, n-gram just makes up its limitation at the expense of corpus size because its detection mechanism for flexible patterns has no radical change.

5 Conclusion

In this article, our LCS-based approach is inspired by gene sequence alignment. In a new view, we reconsider MWE extraction task. These two tasks coincide with each other in pattern recognition. Some new phenomena in natural language are also observed. For example, we improve MWE mining result by EBP detection. Comparisons with variant n-gram approaches, which are the leading approaches, are performed for verifying the effectiveness of our approach. Although LCS approach results in better extraction model, a lot of improvements for more robust model are still needed.

Each innovation presented here only opens the way for more research. Some established theories between Computational Linguistics and Bioinformatics can be shared in a broader way.

6 Acknowledgements

The authors would like to thank three anonymous reviewers for their careful reading and helpful suggestions. This work is supported by National Natural Science Foundation of China (NSFC) (No.60496326) and 863 project of China (No.2001AA114210-11). Our thanks also go to Yushi Xu and Hui Liu for their coding and technical support.

References

Arantza Casillas, Raquel Martłnez , 2002. Aligning Multiword Terms Using a Hybrid Approach. Lecture Notes in Computer Science 2276: The 3rd International Conference of Computational Linguistics and Intelligent Text Processing.

Argamon, Shlomo, Ido Dagan and Yuval Krymolowski, 1999. A memory based approach to learning shallow natural language patterns. Journal of Experimental and Theoretical AI. 11, 369-390.

Beatrice Daille, 2003. Terminology Mining. Lecture Notes in Computer Science 2700: Extraction in the Web Era.

Changning Huang, Endong Xun, Zhou Ming, 2000. A Unified Statistical Model for the Identification of English BaseNP. The 38th Annual Meeting of the Association for Computational Linguistics.

Daniel S. Hirschberg, 1977. Algorithms for the Longest Common Subsequence Problem, Journal of the ACM, 24(4), 664-675.

Diana Binnenpoorte, Catia Cucchiarini, Lou Boves and Helmer Strik,2005. Multiword expressions in spoken language: An exploratory study on pronunciation variation. Computer Speech and Language,19(4):433-449

Hans Peter Lenhof, Burkhard Morgenstern, Knut Reinert, 1999. An exact solution for the segment-to-segment multiple sequence alignment problem. Bioinformatics. 15(3): 203-210.

Ivan A. Sag, Timothy Baldwin, Francis Bond, Ann A. Copestake, Dan Flickinger, 2002. Multiword Expressions: A Pain in the Neck for NLP. Lecture Notes in Computer Science 2276: The 3rd International Conference of Computational Linguistics and Intelligent Text Processing.

Jakob. H. Havgaard, R. Lyngs, G .D. Stormo and J. Gorodkin, 2005. Pairwise local structural alignment of RNA sequences with sequence similarity less than 40 percernt. Bioinfomrmatics. 21(9), 1815-1824.

Joaquim Ferreira da Silva, Gael Dias, Sylvie Guillore, Jose Gabriel Pereira Lopes, 1999. Using LocalMaxs Algorithm for the Extraction of Contiguous and Non-contiguous Multiword Lexical Units. The 9th Portuguese Conference on Artificial Intelligence.

Jordi Vivaldi, Llułs Marquez, Horacio Rodrłguez, 2001. Improving Term Extraction by System Combination Using Boosting. Lecture Notes in Computer Science 2167: The 12th European Conference on Machine Learning.

Kalliopi Zervanou and John McNaught, 2000. A Term-Based Methodology for Template Creation in Information Extraction. Lecture Notes in Computer Science 1835: Natural Language Processing.

Katerina Frantzi, Sophia Ananiadou, Hideki Mima, 2000. Automatic recognition of multi-word terms: the C-value/NC-value method. Int J Digit Libr. 3(2), 115C130.

Knut Reinert, Jens Stoye, Torsten Will, 2000. An iterative method for faster sum-of-pairs multiple sequence alignment. Bioinformatics. 16(9): 808-814.

Makoto Nagao, Shinsuke Mori, 1994. A New Method of N-gram Statistics for Large Number of n and Automatic Extraction of Words and Phrases from Large Text Data of Japanese. The 15th International Conference on Computational Linguistics.

Manning,C.D.,H.,Schutze,1999.Foundations of statistical natural language processing. MIT Press.

Marcus A. Zachariah, Gavin E. Crooks, Stephen R. Holbrook, Steven E. Brenner, 2005. A Generalized Affine Gap Model Significantly Improves Protein Sequence Alignment Accuracy. PROTEINS: Structure, Function, and Bioinformatics. 58(2), 329 - 338

Michael. Sammeth, B. Morgenstern, and J. Stoye, 2003. Divide-and-conquer multiple alignment with segment-based constraints. Bioinformatics. 19(2), 189-195.

Mike Paterson, Vlado Dancik ,1994. Longest Common Subsequences. Mathematical Foundations of Computer Science.

Pascale Fung, Kathleen Mckeown, 1997. A Technical Word and Term Translation Aid Using Noisy Parallel Corpora across Language Groups. Machine Translation. 12, 53C87.

Paul Deane, 2005. A Nonparametric Method for Extraction of Candidate Phrasal Terms. The 43rd Annual Meeting of the Association for Computational Linguistics.

Robertson, A.M. and Willett, P., 1998. Applications of n-grams in textual information systems. Journal of Documentation, 54(1), 48-69.

Satanjeev Banerjee, Ted Pedersen, 2003. The Design, Implementation, and Use of the Ngram Statistics Package. Lecture Notes in Computer Science 2588: The 4th International Conference of Computational Linguistics and Intelligent Text Processing.

Smith, T.F., Waterman, M.S., 1981. Identification of common molecular subsequences. J. Molecular Biology. 147(1), 195-197.

Stefan Diaconescu, 2004. Multiword Expression Translation Using Generative Dependency Grammar. Lecture Notes in Computer Science 3230: Advances in Natural Language Processing.

Suleiman H. Mustafa, 2004. Character contiguity in N-gram-based word matching: the case for Arabic text searching. Information Processing and Management. 41(4), 819-827.

Taneli Mielikainen, 2003. Frequency-Based Views to Pattern Collections. IFIP/SIAM Workshop on Discrete Mathematics and Data Mining.

Violeta Seretan, Luka Nerima, Eric Wehrl, 2003. Extraction of Multi-Word Collocations Using Syntactic Bigram Composition. International Conference on Recent Advances in NLP.