Propositional Term Extraction over Short Text using Word Cohesiveness and Conditional Random Fields with Multi-Level Features

張如瑩　Ru-Yng Chang
國立成功大學資訊工程學系
Department of Computer Science and Information Engineering
National Cheng Kung University
ruyng@csie.ncku.edu.tw

吳宗憲　Chung-Hsien Wu
國立成功大學資訊工程學系
Department of Computer Science and Information Engineering
National Cheng Kung University
chwu@csie.ncku.edu.tw

Abstract

Propositional terms in a research abstract (RA) generally convey the most important information for readers to quickly glean the contribution of a research article. This paper considers propositional term extraction from RAs as a sequence labeling task using the IOB (Inside, Outside, Beginning) encoding scheme. In this study, conditional random fields (CRFs) are used to initially detect the propositional terms, and the combined association measure (CAM) is applied to further adjust the term boundaries. This method can extract beyond simply NP-based propositional terms by combining multi-level features and inner lexical cohesion. Experimental results show that CRFs can significantly increase the recall rate of imperfect boundary term extraction and the CAM can further effectively improve the term boundaries.

摘要

命題術語(Propositional Term)表達文章中重要概念且引導讀者文章脈絡之發展。這篇論文以學術論文摘要為實驗對象進行命題術語擷取，研究中整合條件隨機域(Conditional Random Fields, CRFs) 以及結合聯繫測量(Combined Association Measure, CAM) 兩種方法，考量詞彙內部凝聚力和文脈兩大類訊息，截取出的命題術語不再侷限於名詞片語型態，且可由單詞或多詞所構成。在命題術語擷取的過程中，將其視為一種序列資料標籤的任務，並利用 IOB 編碼方式識別命題述語的邊界，CRF 考量多層次構成命題述語的特徵，負責初步命題術語偵測，再利用 CAM 計算詞彙凝聚力，藉以加強確認命題術語詞彙的邊界。實驗結果顯示 ，本研究所提出的方法比以往述語偵測方法在效能上有明顯增進，其中，CRF 明顯增進非完美術語詞彙邊界辨識(Imperfect hits)的召回率，而 CAM 則有效修正術語詞彙邊界。

Keywords: Propositional Term Extraction, Conditional Random Fields, Combined Association Measure, Multi-Level Feature

關鍵詞：命題述語擷取，條件隨機域，結合聯繫測量，多層次特徵

1. Introduction

Researchers generally review Research Abstracts (RAs) to quickly track recent research trends. However, many non-native speakers experience difficulties in writing and reading RAs [1]. The author-defined keywords and categories of the research articles currently utilized to provide researchers with access to content guiding information are cursory and general. Therefore, developing a propositional term extraction system is an attempt to exploit the linguistic evidence and other characteristics of RAs to achieve efficient paper comprehension. Other applications of the proposed method contain sentence extension, text generation, and content summarization.

A term is a linguistic representation of a concept with a specific meaning in a particular field. It may be composed of a single word (called a simple term), or several words (a multiword term) [2]. A propositional term is a term that refers to the basic meaning of a sentence (the proposition) and helps to extend or control the development of ideas in a text. The main difference between a term and a propositional term is that a propositional term, which can guide the reader through the flow of the content, is determined by not only syntax or morphology but semantic information. Take RAs to illustrate the difference between a term and a propositional term. Cheng [3] indicted that a science RA is composed of background, manner, attribute, comparison and evaluation concepts. In Figure 1, the terms underlined are the propositional terms which convey the important information of the RA. In the clause “*we present one of the first robust LVCSR systems that use a syllable-level acoustic unit for LVCSR,*” the terms “*LVCSR systems*”, “*syllable-level acoustic unit*” and “*LVCSR*” respectively represent the background, manner and background concepts of the research topic, and can thus be regarded as propositional terms in this RA. The background concepts can be identified by clues from the linguistic context, such as the phrases “*most…LVCSR systems*” and “*in the past decade*”, which indicate the aspects of previous research on LVCSR. For the manner concept, contextual indicators such as the phrases “*present one of…*”, “*that use*” and “*for LVCSR*” express the aspects of the methodology used in the research. Propositional terms may be composed of a variety of word forms and syntactic structures and thus may not only be NP-based, and therefore cannot be extracted by previous NP-based term extraction approaches.

> Most large vocabulary continuous speech recognition (LVCSR) systems in the past decade have used a context-dependent (CD) phone as the fundamental acoustic unit. In this paper, we present one of the first robust LVCSR systems that use a syllable-level acoustic unit for LVCSR on telephone-bandwidth speech. This effort is motivated by the inherent limitations in phone-based approaches-namely the lack of an easy and efficient way for modeling long-term temporal dependencies. A syllable unit spans a longer time frame, typically three phones, thereby offering a more parsimonious framework for modeling pronunciation variation in spontaneous speech. We present encouraging results which show that a syllable-based system exceeds the performance of a comparable triphone system both in terms of word error rate (WER) and complexity. The WER of the best syllable system reported here is 49.1% on a standard SWITCHBOARD evaluation, a small improvement over the triphone system. We also report results on a much smaller recognition task, OGI Alphadigits, which was used to validate some of the benefits syllables offer over triphones. The syllable-based system exceeds the performance of the triphone system by nearly 20%, an impressive accomplishment since the alphadigits application consists mostly of phone-level minimal pair distinctions.

Figure1. A Manually-Tagged Example of Propositional Terms in an RA

In the past, there were three main approaches to term extraction: linguistic [4], statistical [5, 6], and C/NC-value based [7,8] hybrid approaches. Most previous approaches can only achieve a good performance on a test article composed of a relatively large amount of words. Without the use of large amount of words, this study proposes a method for extracting and weighting single- and multi-word propositional terms of varying syntactic structures.

2. System Design and Development

This research extracts the propositional terms beyond simply the NP-based propositional terms from the abstract of technical papers and then regards propositional term extraction as a sequence labeling task. To this end, this approach employs an IOB (Inside, Outside, Beginning) encoding scheme [9] to specify the propositional term boundaries, and conditional random fields (CRFs) [10] to combine arbitrary observation features to find the globally optimal term boundaries. The combined association measure (CAM) [11] is further adopted to modify the propositional term boundaries. In other words, this research not only considers the multi-level contextual information of an RA (such as word statistics, tense, morphology, syntax, semantics, sentence structure, and cue words) but also computes the lexical cohesion of word sequences to determine whether or not a propositional term is formed, since contextual information and lexical cohesion are two major factors for propositional term generation.

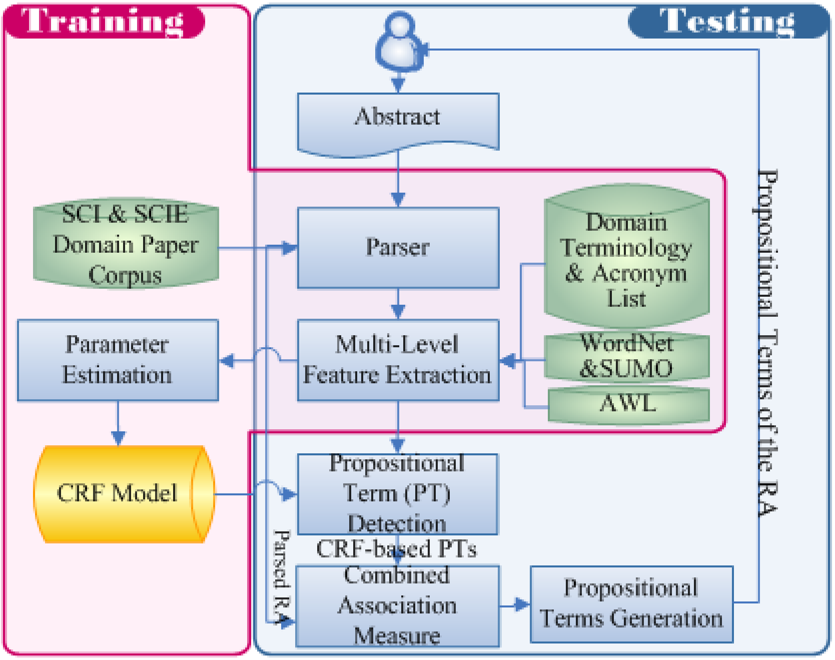
Training
Testing
Abstract
SCI & SCIE Domain Paper Corpus
Parser
Domain Terminology & Acronym List
Parameter Estimation
Multi-Level Feature Extraction
WordNet &SUMO
AWL
CRF Model
Propositional Term (PT) Detection
CRF-based PTs
Parsed RA
Combined Association Measure
Propositional Terms Generation
Propositional Terms of the RA

Figure 2. The System Framework of Propositional Term Extraction

The system framework essentially consists of a training phase and a test phase. In the training phase, the multi-level features were extracted from specific domain papers which were gathered from the SCI (Science Citation Index)-indexed and SCIE (Science Citation Index Expanded)-indexed databases. The specific domain papers are annotated by experts and then parsed. The feature extraction module collects statistical, syntactic, semantic and morphological level global and local features, and the parameter estimation module calculates conditional probabilities and optimal weights. The propositional term detection CRF model was built with feature extraction module and the parameter estimation module. During the test phase users can input an RA and obtain system feedback, i.e. the propositional terms of the RA. When the CRF model produces the preliminary candidate propositional terms, the propositional term generation module utilizes the combined association measure (CAM) to adjust the propositional term boundaries. The system framework proposed in this paper for RA propositional term extraction is shown in Figure 2. A more detailed discussion is presented in the following subsections.

2.1. Assisted Resource

In order to produce different levels of information and further assist feature extraction in the training and test phases, several resources were employed. This study chooses the ACM Computing Classification System (ACM CSS) [12] to serve as the domain terminology list for propositional term extraction from computer science RAs. The ACM CSS provides important subject descriptors for computer science, and was developed by the Association for Computing Machinery. The ACM CSS also provides a list of Implicit Subject Descriptors, which includes names of languages, people, and products in the field of computing. A mapping database, derived from WordNet (http://wordnet.princeton.edu/) and SUMO (Suggested Upper Merged Ontology) (http://ontology.teknowledge.com/) [13], supplies the semantic concept information of each word and the hierarchical concept information from the ontology. The AWL (Academic Words List) (http://www.vuw.ac.nz/lals/research/awl/) [14] is an academic word list containing 570 word families whose words are selected from different subjects. The syntactic level information of the RAs was obtained using Charniak's parser [15], which is a "maximum-entropy inspired" probabilistic generative model parser for English.

2.2. Conditional Random Fields (CRFs)

For this research goal, given a word sequence $W=\{w_1,w_2,...,w_n\}$, the most likely propositional term label sequence $S=\{s_1,s_2,...,s_n\}$ in the CRF framework with the set of weights Ψ can be obtained from the following equation.

$$\hat{S}=\arg\max_S P_\Psi\left(S\,|\,W\right) \qquad (1)$$

A CRF is a conditional probability sequence as well as an undirected graphical model which defines a conditional distribution over the entire label sequence given the observation sequence. Unlike Maximum Entropy Markov Models (MEMMs), CRFs use an exponential model for the joint probability of the whole label sequence given the observation to solve the label bias problem. CRFs also have a conditional nature and model the real-world data depending on non-independent and interacting features of the observation sequence. A CRF allows the combination of overlapping, arbitrary and agglomerative observation features from both the past and future. The propositional terms extracted by CRFs are not restricted by syntactic variations or multiword forms and the global optimum is generated from different global and local contributor types.

The CRF consists of the observed input word sequence $W=\{w_1,w_2,...,w_n\}$ and label state sequence $S=\{s_1,s_2,...,s_n\}$ such that the expansion joint probability of a state label sequence given an observation word sequence can be written as

$$P\left(S\,|\,W\right)=\frac{1}{Z_0}\exp\left(\sum_t\sum_k\lambda_k f_k\left(s_{t-1},s_t,W\right)+\sum_t\sum_k\mu_k g_k\left(s_t,W\right)\right) \qquad (2)$$

where $f_k\left(s_{t-1},s_t,W\right)$ are the transition features of the global observation sequence and the states at positions t and t-1 in the corresponding state sequence, and $g_k\left(s_t,W\right)$ is a state feature function of the label at position t and the observation sequence. Let λ_k be the weight of each f_k, μ_k be the weight of g_k and $\frac{1}{Z_0}$ be a normalization factor over all state sequences, where $Z_0 = \sum_S \exp\left(\sum_t \sum_k \lambda_k f_k\left(s_{t-1}, s_t, W\right) + \sum_t \sum_k \mu_k g_k\left(s_t, W\right) \right)$.

The set of weights in a CRF model, $\Psi = \left(\lambda_k, \mu_k\right)$, is usually estimated by maximizing the conditional log-likelihood of the labeled sequences in the training data $D = \left\{S^{(i)}, W^{(i)}\right\}_{i=1}^{n}$. (Equation (3)) For fast training, parameter estimation was based on L-BFGS (the limited-memory BFGS) algorithm, a quasi-Newton algorithm for large scale numerical optimization problems [16]. The L-BFGS had proved [17] that converges significantly faster than Improved Iterative Scaling (IIS) and General Iterative Scaling (GIS).

$$L_\Psi = \sum_{i=1\ldots N} \log\left(P_\Psi\left(S^{(i)} \mid W^{(i)}\right)\right) \tag{3}$$

After the CRF model is trained to maximize the conditional log-likelihood of a given training set P(S|W), the test phase finds the most likely sequence using the combination of forward Viterbi and backward A* search [18]. The forward Viterbi search makes the labeling task more efficient and the backward A* search finds the n-best probable labels.

2.3. Multi-Level Features

According to the properties of propositional term generation and the characteristics of the CRF feature function, this paper adopted local and global features which consider statistical, syntactic, semantic, morphological, and structural level information. In the CRF model, the features used were binary and were formed by instantiating templates, and the maximum entropy principle was provided for choosing the potential functions. Equation (4) shows an example of a feature function, which was set to 1 when the word was found in the rare words list (RW).

$$g_{s,w_1,w_2,\ldots,w_n}\left(s_t, w_1^n\right) = \begin{cases} 1, \text{ if } \mathrm{s_t} = s \cap isRW\left(\mathrm{W_t}\right) \\ 0, \text{ otherwise} \end{cases} \tag{4}$$

2.3.1. Local Feature

(1). Morphological Level:

Scientific terminology often ends with similar words, e.g. "*algorithm*" or "*model*", or is represented by connected words (CW) expressed with hyphenation, quotation marks or brackets. ACMCSS represents entries in the ACM Computing Classification System (ACM CSS). The last word of every entry in the ACM CSS (ACMCSSAff) satisfies the condition that it is a commonly occurring last word in scientific terminology. The existing propositional terms of the training data were the seeds of multiword terms (MTSeed).

Words identified as acronyms were stored as useful features, consisting of IsNenadic, IsISD, and IsUC. IsNenadic was defined using the methodology of Nenadić, Spasić and Ananiadou [19] to acquire possible acronyms of a word sequence that was extracted by the C/NC value method. IsISD refers to the list of Implicit Subject Descriptors in the ACM CCS and IsUC signifies that all characters of the word were uppercase

(2). Semantic Level:

MeasureConcept infers that the word was found under SUMO's

"UNITS-OF-MEASURE" concept subclass and SeedConcept denotes that the concept of the word corresponded to the concept of a propositional term in the training data.

(3). Frequency Level:

A high frequency word list (HF) was generated from the top 5 percent of words in the training data. A special words list (SW) consists of the out-of-vocabulary and rare words. Out-of-vocabulary words are those words that do not exist in WordNet. Rare words are words not appearing in the AWL or which appear in less than 5 different abstracts.

(4). Syntactic Level:

This feature was set to 1 if the syntactic pattern of the word sequence matched the regular expression "*(NP)*(preposition)?(NP)**" (SynPattern), or matched the terms in the training data (SeedSynPattern). SyntaxCon means that concordances of ACMCSSAff or ACMCSSAffSyn (ACMCSSAff synonyms) used the keyword in context to find the syntactic frame in the training data. If the part-of-speech (POS) of the word was a cardinal number, then this feature CDPOS was set to 1.

(5). Statistical and Syntactic Level:

This research used the CRF model to filter terms extracted by the C/NC value approach with no frequency threshold

2.3.2. Global Feature

(1). Cue word:

KeyWord infers that the word sequence matched one of the user's keywords or one word of the user's title. IsTransW and IsCV represent that a word was found in an NP after TransW or CV respectively. TransW indicates summative and enumerative transitional words, such as "*in summary*", "*to conclude*", "*then*", "*moreover*", and "*therefore*", and CV refers to words under SUMO's "*communication*" concepts, such as "*propose*", "*argue*", "*attempt*" and so on.

(2). Tense:

If the first sentence of the RA is in the past tense and contains an NP, then the word sequence of that NP was used as a useful feature PastNP. This is because the first sentence often impresses upon the reader the shortest possible relevant characterization of the paper, and the use of past tense emphasizes the importance of the statement.

(3). Sentence structure:

Phrases in a parallel structure sentence refers to the phrases appearing in a sentence structure such as Phrase, Phrase, or (and) Phrase, and implies that the same pattern of words represents the same concept. ParallelStruct indicates that the word was part of a phrase in a parallel structure.

2.4. Word Cohesiveness Measure

By calculating the cohesiveness of words, the combined association measure (CAM) can assist in further enhancing and editing the CRF-based propositional term boundaries for achieving a perfect boundary of propositional terms. CAM extracts the most relevant word sequence by combining endogenous linguistic statistical information, including word form sequence and its POS sequence. CAM is a variant of normalized expectation (NE) and mutual expectation (ME) methods.

To characterize the degree of cohesiveness of a sequence of textual units, NE evaluates the average cost of loss for a component in a potential word sequence. NE is defined in Equation (5) where the function c(·) means the count of any potential word sequence. An example of NE is shown in Equation (6).

$$NE\left(\left[w_1 ... w_i ... w_n\right]\right) = \frac{C\left(\left[w_1 ... w_i ... w_n\right]\right)}{\frac{1}{n}\left(C\left(\left[w_1 ... w_i ... w_n\right]\right) + \sum_{i=2}^{n} C\left(\left[w_1 ... \hat{w}_i ... w_n\right]\right)\right)} \tag{5}$$

$$\begin{aligned} & NE\left(\left[\text{large vocabulary continuous speech recognition}\right]\right) \\ & = \frac{C\left(\left[\text{large vocabulary continuous speech recognition}\right]\right)}{\frac{1}{5}\left(\begin{array}{l} C\left(\left[\text{large vocabulary continuous speech recognition}\right]\right) \\ +C\left(\left[\text{large continuous speech recognition}\right]\right) \\ +C\left(\left[\text{large vocabulary speech recognition}\right]\right) \\ +C\left(\left[\text{large vocabulary continuous recognition}\right]\right) \\ +C\left(\left[\text{large vocabulary continuous speech}\right]\right) \end{array}\right)} \end{aligned} \tag{6}$$

Based on NE and relative frequency, the ME of any potential word sequence is defined as Equation (7), where function P(·) represents the relative frequency.

$$ME\left(\left[w_1 ... w_i ... w_n\right]\right) = P\left(\left[w_1 ... w_i ... w_n\right]\right) \times NE\left(\left[w_1 ... w_i ... w_n\right]\right) \tag{7}$$

CAM considers that the global degree of cohesiveness of any word sequence is evaluated by integrating the strength in a word sequence and the interdependence of its POS. Thus CAM evaluates the cohesiveness of a word sequence by the combination of its own ME and the ME of its associated POS sequence. In Equation (8), CAM integrates the ME of word form sequence $\left[w_1 ... w_i ... w_n\right]$ and its POS $\left[p_1 ... p_i ... p_n\right]$. Let α be a weight between 0 and 1, which determines the degree of the effect of POS or word sequence in the word cohesiveness measure.

$$CAM\left(\left[w_1 ... w_i ... w_n\right]\right) = ME\left(\left[w_1 ... w_i ... w_n\right]\right)^{\alpha} \times ME\left(\left[p_1 ... p_i ... p_n\right]\right)^{1-\alpha} \tag{8}$$

This paper uses a sliding window moving in a frame and compares the CAM value of neighboring word sequences to determine the optimal propositional term boundary. Most lexical relations associate words distributed by the five neighboring words [20]. Therefore this paper only calculates the CAM value of the three words to the right and the three words to the left of the CRF-based terms. Figure 3 represents an illustration for the CAM computation that was fixed in the [(2*3) + length(CRF-Based term)] frame size with a sliding window. When the window starts a forward or backward move in the frame, the three marginal words of a term are the natural components of the window. As the word number of the CRF term is less than three words, the initial sliding windows size is equal to the word number of the term.

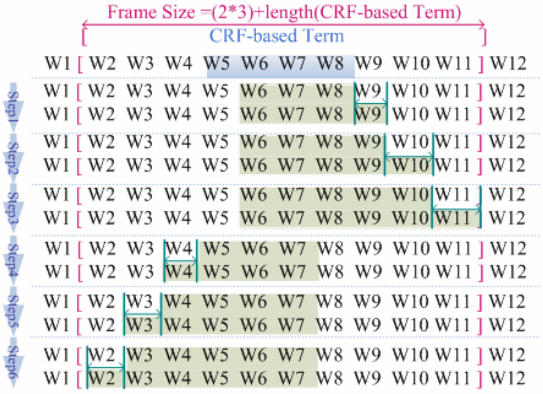
Frame Size =(2*3)+length(CRF-based Term)
CRF-based Term
W1 [W2 W3 W4 W5 W6 W7 W8 W9 W10 W11] W12
Step1
Step2
Step3
Step4
Step5
Step6

Figure 3. An Illustration for the CAM Computation Steps

To find the optimal propositional term boundary, this study calculates the local maximum CAM value by using the Modified CamLocalMax Algorithm. The principle of the original algorithm [21] is to infer the word sequence as a multiword unit if the CAM value is higher than or equal to the CAM value of all its sub-group of (n-1) words and if the CAM value is higher than the CAM value of all its super-group of (n+1) words. In the Modified CamLocalMax Algorithm, when the CAM value of the combination of CRF-based single word propositional terms and its immediate neighbor word is higher than the average of the CAM value of bi-gram propositional terms in the training data, the components of the CRF-based single word propositional terms are turned into a bi-gram propositional term. The complete Modified CamLocalMax Algorithm is shown in the following, where *cam* means the combined association measure, size(·) returns the number of words of a possible propositional term, *M* represents a possible propositional term, Ω_{n+1} denotes the set of all the possible (n+1)grams containing *M*, Ω_{n-1} denotes the set of all the possible (n-1)grams contained in *M*, and bi-term typifies bi-gram propositional terms in the training data.

```
Input: M, a possible propositional term, $\forall y \in \Omega_{n+1}$, the set of all the possible (n+1)grams
containing M, $\forall x \in \Omega_{n-1}$, the set of all the possible (n-1)grams contained in M
Output: CT={$ct_1$,$ct_2$,…$ct_n$}, a CRF+CAM-based propositional term set
If    (size(M)=2 and    cam(M) > cam(y))
   or ( size(M)>2 and    cam(M) $\geqq$ cam(x)    and cam(M) >cam(y) )
   or ( size(M)=1 and cam(bi-gram) $\leqq$ cam(M) )
End if
Return ct
```

2.5. Propositional Term Generation Algorithm

The Propositional Term Generation algorithm utilizes the CRF model to generate a CRF-based propositional term set T={t_1,t_2,…t_n} and calculates the CAM value to produce a CRF+CAM-based propositional term set CT={ct_1,ct_2,…ct_n}. The detailed processes of the Propositional Term Generation algorithm are as follows

```
$t_n^k$ : the word form sequence from the first word 1 to last word k of CRF-based propositional term $t_n$
Input: Word sequence $W_1^n$
Output: T={$t_1$,$t_2$,…$t_n$}, a CRF-based propositional term set and, CT={$ct_1$,$ct_2$,…$ct_n$}, a CRF+CAM-based
        propositional term set
Input $W_1^n$ to generate T={$t_1$,$t_2$,…$t_n$} by CRF
For all $t_j \in T$
    For a=0 to a =2 Step 1
```

```
        $ct_j$=Modified_CamLocalMax($t_j^{k+a}, t_j^{k+a-1}, t_j^{k+a+1}$)
        $CT \leftarrow CT \cup ct$
        End for
        If $t_j \notin CT$ Then
            For a=0 to a =-2 Step -1
                $ct_j$=Modified_CamLocalMax($t_j^{1+a}, t_j^{1+a-1}, t_j^{1+a+1}$)
                $CT \leftarrow CT \cup ct_j$
            End for
        End if
End for
Return T, CT
```

2.6. Encoding Schema

The IOB encoding scheme was adopted to label the words, where I represents words Inside the propositional term, O marks words Outside the propositional term, and B denotes the Beginning of a propositional term. It should be noted that here the B tag differs slightly from Ramshaw and Marcus's definition, which marks the left-most component of a baseNP for discriminating recursive NPs. Figure 4 shows an example of the IOB encoding scheme that specifies the B, I, and O labels for the sentence fragment "*The syllable-based system exceeds the performance of the triphone system by…*". An advantage of this encoding scheme is that it can avoid the problem of ambiguous propositional term boundaries, since IOB tags can identify the boundaries of immediate neighbor propositional terms, whereas binary-based encoding schemes cannot. In Figure 4, "*syllable-based system*", and "*exceeds*" are individual and immediate neighbor propositional terms distinguished by B tags.

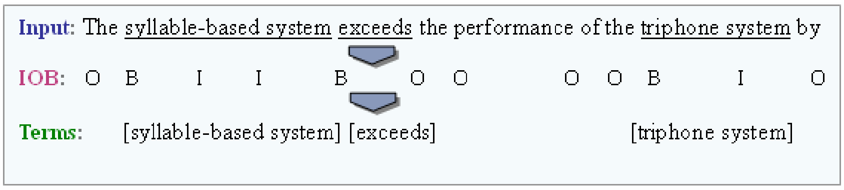
Input: The syllable-based system exceeds the performance of the triphone system by
IOB: O B I I B O O O O B I O
Terms: [syllable-based system] [exceeds] [triphone system]

Figure 4. An Example of the IOB Encoding Scheme

3. Evaluation

3.1. Experimental Setup

To facilitate the development and evaluation of the propositional term extraction method, experts manually annotated 260 research abstracts, including speech, language, and multimedia information processing journal papers from SCI and SCIE-indexed databases. In all, there were 109, 72, and 79 annotated research abstracts in the fields of speech, language, and multimedia information processing, respectively. At run time, 90% of the RAs were allocated as the training data and the remaining 10% were reserved as the test data for all evaluation.

In system implementation, the CRF++: Yet Another CRF toolkit 0.44 [22] was adopted. The training parameters were chosen using ten-fold cross-validation on each experiment.

The proposed system was compared with three baseline systems. The first was the C/NC-value algorithm with no frequency threshold, because the C/NC-value algorithm is a hybrid methodology and its historical result is better than the linguistic and statistical approaches. The second baseline system proposed by Nenadić et al. [8] is a variant of the

C/NC-value algorithm enriched by morphological and structural variants. The final baseline system is a linguistic approach proposed by Ananiadou [4]. That study, however, made no comparisons with statistical approaches which are suitable for a document containing a large amount of words.

To evaluate the performance in this study, two hit types for propositional term extraction: perfect and imperfect [23] are employed. A perfect hit means that the boundaries of a term's maximal term form conform to the boundaries assigned by the automatic propositional term extraction. An imperfect hit means that the boundaries assigned by the automatic propositional term extraction do not conform to the boundaries of a term's maximal term form but include at least one word belonging to a term's maximal term form. Taking the word sequence "*large vocabulary continuous speech recognition*" as an example, when the system detects that "*vocabulary continuous speech recognition*" is a propositional term, it then becomes an imperfect hit. There is only one perfect hit condition where "*large vocabulary continuous speech recognition*" is recognized. The metrics of recall and precision were also used to measure the perfect and imperfect hits. The definition of recall and precision of perfect hits and imperfect hits are shown in Equation (9) and Equation (10). Thus, our system is evaluated with respect to the accuracies of propositional term detection and propositional term boundary detection. That is, our motivation for propositional term extraction was to provide CRF and CRF+CAM for accurate detection of propositional terms and the improvement of the detected propositional term boundaries.

$$\text{Recall}=\text{Hits Perfect (or Imperfect)} \big/ \text{Target Termforms} \tag{9}$$

$$\text{Precision}=\text{Hits Perfect (or Imperfect)} \big/ \text{Extracted Termforms} \tag{10}$$

3.2. Experimental Results

This study evaluated empirically two aspects of our research for different purposes. First, the performance of propositional term extraction for CRF-based and CRF+CAM-based propositional term sets on different data was measured. Second, the impact of different level features for propositional term extraction using CRF was evaluated.

Evaluation of Different Methods

Table 1. The Performance of Imperfect Hits on Different Data

Method	R	P	F	R	P	F
	All Data			Language Data		
CRF Inside Testing	93.2	94.5	93.9	96.7	98.1	97.4
CRF +CAM Inside Testing	96.6	96.0	96.3	98.4	99.6	99.0
CRF Outside Testing	77.1	74.1	75.6	78.6	76.3	77.4
CRF +CAM Outside Testing	82.6	82.5	82.6	85.8	88.8	87.2
C/NC Value	53.4	65.3	58.8	48.1	53.3	50.6
Ananiadou	51.3	70.0	59.2	52.4	68.4	59.3
Nenadić et al.	58.0	72.3	64.4	60.1	69.0	64.3
	Speech Data			Multimedia Data		
CRF Inside Testing	96.6	99.0	98.2	98.0	99.2	98.6
CRF +CAM Inside Testing	97.5	99.0	99.4	98.6	99.3	99.0
CRF Outside Testing	74.9	76.1	74.3	61.2	65.0	63.1
CRF +CAM Outside Testing	82.6	83.9	84.2	65.4	71.2	68.2
C/NC Value	53.5	79.0	62.7	67.7	53.2	59.6
Ananiadou	53.1	68.4	59.8	65.4	60.0	62.6

Nenadić et al.	59.6	72.2	65.3	68.9	55.2	61.3

Table 1 lists the recall rate, the precision rate and F-score of propositional term extraction for imperfect hits of different domain data. In each case, the recall and precision of imperfect hits using CRF inside testing was greater than 93%. The CRF outside test achieved approximately 73% average recall and 73% average precision for imperfect hits, and the CAM approach improved the original performance of recall and precision for imperfect hits. The C/NC-value approach achieved approximately 56% average recall and 63% average precision for imperfect hits. The performance of Ananiadou's approach was about 56% average recall and 67% average precision for imperfect hits. Another baseline, the approach of Nenadić, Ananiadou and McNaught, obtained approximately 62% average recall and 67% average precision for imperfect hits.

Table 2. The Performance of Perfect Hits on Different Data

Method	R	P	F	R	P	F
	All Data			Language Data		
CRF Inside Testing	66.5	66.2	66.3	66.4	67.5	67.0
CRF +CAM Inside Testing	69.0	68.6	68.8	69.4	69.9	69.6
CRF Outside Testing	39.8	42.2	41.9	43.2	37.3	40.0
CRF +CAM Outside Testing	43.5	49.2	46.2	45.3	45.4	45.3
C/NC Value	27.6	37.8	31.9	28.9	29.1	29.0
Ananiadou	26.3	37.9	31.1	31.3	37.7	34.2
Nenadić et al.	30.2	41.0	34.8	31.2	40.9	35.4
	Speech Data			Multimedia Data		
CRF Inside Testing	62.3	61.0	61.7	70.9	70.3	70.6
CRF +CAM Inside Testing	69.6	67.9	68.7	73.1	70.3	71.6
CRF Outside Testing	36.9	41.6	39.1	42.1	42.5	42.3
CRF +CAM Outside Testing	42.8	48.9	45.6	45.6	45.0	44.3
C/NC Value	29.0	40.0	33.6	34.6	29.9	32.1
Ananiadou	27.4	37.7	31.7	29.3	38.0	33.1
Nenadić et al.	30.0	38.6	33.7	35.3	37.6	35.3

Table 2 summarizes the recall rates, precision rates and F-score of propositional term extraction for perfect hits of data from different domains. The CRF inside test achieved approximately 67% average recall and 66% average precision on perfect hits, but the CRF outside test did not perform as well. However, the CAM approach still achieved an increase of 1%-7% for perfect hits. The C/NC-value approach obtained approximately 30% average recall and 34% average precision for perfect hits. Ananiadou's approach achieved approximately 29% average recall and 38% average precision for perfect hits. The performance of Nenadić, Ananiadou and McNaught's approach was about 32% average recall and 40% average precision for perfect hits.

The results show that the C/NC-value does not demonstrate a significant change over different fields, except for the multimedia field, which had slightly better recall rate. The main reasons for errors produced by C/NC-value were propositional terms that were single words or acronyms, propositional terms that were not NP-based, or propositional terms that consisted of more than four words.

Ananiadou's approach was based on a morphological analyzer and combination rules for the different levels of word forms. Experimental results showed that this approach is still unable to deal with single words or acronyms, and propositional terms that are not NP-based.

Nenadić et al.'s approach considered local morphological and syntactical variants using C value to determine the propositional terms. This approach had slightly better performance than the C/NC value methodology. Acronyms were included in the propositional term candidates but were filtered by frequency, as they often appear only a few times. This approach also ignored single words, and propositional terms that were not NP-based. Furthermore, none of these three baseline systems are suitable for handling special symbols.

For CRF inside testing, both the precision and recall rates were significantly better for imperfect hits, but the precision and recall rates were reduced by about 30% for perfect hits in most RAs. Due to insufficient training data, CRF no longer achieved outstanding results. In particular, the large variability and abstract description of the multimedia field RAs led to huge differences between measures. For example, in the sentence "*For surfaces with varying material properties, a full segmentation into different material types is also computed*", "*full segmentation into different material types*" is a propositional term that it isn't concretely specified as a method. CRF achieved a better result in recall rate, but failed on propositional term boundary detection, unlike the C/NC-value approach.

The CAM approach effectively enhanced propositional term boundary detection by calculating word cohesiveness, except in the case of multimedia data. The CAM approach couldn't achieve similar performance for the multimedia data as a result of the longer word count of terms that differ from the data of other fields. However, the CAM approach performed best with α equal to 0.4, which demonstrates that the POS provided a little more contribution for multiword term construction. The CAM approach not only considered the POS sequence but also the word sequence, therefore the results are a little better for speech data, which is the biggest part of the training data (SCI and SCIE-indexed databases).

The above results show that the CRF approach exhibited impressive improvements in propositional term detection. The major reason for false positives was that the amount of the data was not enough to construct the optimal model. Experimental results revealed that the CAM is sufficiently efficient for propositional term boundary enhancement but the longer word count of propositional terms were excluded.

Evaluation of Different Level Features

In order to assess the impact of different level features on the extraction method, this paper also carried out an evaluation on the performance when different level features were omitted. Table 3 presents the performance of CRF when omitting different level features for imperfect hits and the symbol "-" denoted the test without a level feature. For all data, the recall rate was reduced by approximately 1%- 5% and the precision rate was reduced by approximately 2%- 6% in inside testing result. In all data outside testing, the recall rate was reduced by 2%-10% and the precision rate was reduced by 1%-5%. The recall and precision for speech data retained similar results from semantic level features, but showed little impact from other local features. For language data, without morphological, syntactic, frequency, and syntactic & statistical level features the performance was slightly worse than the original result and without semantic level features the original performance was preserved. The performance for multimedia data was affected greatly by semantic level features. A slight improvement without morphological, and syntactic & statistical level features and similar results were obtained when frequency and syntactic level features were omitted.

Table 3. The Performance of CRF Excepting Different Level Features for Imperfect Hits

Data Type / Testing Type	All		Speech		Language		Multimedia	
	R	P	R	P	R	P	R	P
Inside -Frequency Features	92	92	94	97	95	97	98	98
Inside -Morphological Features	88	90	92	96	93	96	97	97
Inside -Syntactic Features	90	89	94	96	95	97	97	98
Inside -Semantic Features	92	92	96	98	97	98	95	97
Inside -Syntactic & Statistical Features	90	93	93	95	95	96	96	98
Inside Testing	93	95	97	99	97	98	98	99
Outside -Frequency Features	74	73	71	73	76	74	60	65

Outside -Morphological Features	71	71	59	69	70	68	58	65
Outside -Syntactic Features	67	69	60	71	71	71	59	64
Outside -Semantic Features	75	75	75	76	78	76	41	60
Outside -Syntactic &Statistical Features	71	73	67	71	70	70	55	65
Outside Testing	77	74	75	76	79	76	61	65

In Table 4, it can be noticed that the omission of any single level features results in a deterioration in the performance of perfect hits. Removing the syntactic level features had the most pronounced effect on performance for all, speech and language data, while removing the semantic level features had the least effect on performance for all, speech and language data. According to the experimental results, the use of the frequency features did not result in any significant performance improvement for the multimedia data, and the use of the syntactic and syntactic & statistical level features did not result in any performance improvement for the multimedia data. Removing the semantic level features had the greatest effect on the performance for the multimedia data.

Table 4. The Performance of CRF without Different Level Features for Perfect Hits

Data Type	All		Speech		Language		Multimedia	
Testing Type	R	P	R	P	R	P	R	P
Inside -Frequency Features	63	60	56	55	61	64	60	60
Inside -Morphological Features	61	61	57	54	61	64	70	68
Inside -Syntactic Features	60	60	55	57	63	65	68	67
Inside -Semantic Features	65	62	59	60	66	69	62	62
Inside -Syntactic &Statistical Features	62	61	57	52	62	64	71	68
Inside Testing	67	66	62	61	66	68	71	70
Outside -Frequency Features	36	38	34	35	37	34	40	40
Outside -Morphological Features	33	35	32	36	35	34	40	39
Outside -Syntactic Features	35	36	32	38	37	32	39	40
Outside -Semantic Features	38	40	36	40	41	36	29	31
Outside -Syntactic &Statistical Features	38	39	32	37	35	33	40	40
Outside Testing	40	42	37	42	42	37	42	42

Overall the five different level features were all somewhat effective for propositional term extraction. This suggests that propositional terms are determined by different level feature information which can be effectively used for propositional term extraction. The frequency level features contributed little for propositional term extraction in all and speech data. This may be due to the fact that speech data comprised the main portion of the training data. In the multimedia case, the semantic level features were useful. Although semantic level features may include some useful information, it was still a problem to correctly utilize such information in the different domain data for propositional term extraction. Syntactic and morphological level features obtained the best performance for all, speech and language data. This may be due to the amount of training data in each domain and the various word forms of propositional terms in the multimedia data. The syntactic and statistical level features improved or retained the same performance, which indicates the combined effectiveness of syntactic and statistical information.

3.3. Error Analysis

Table 5 shows the distribution of error types on propositional term extraction for each domain data using outside testing. This study adopts the measure used in [24] to evaluate the error type, where M indicates the condition when the boundary of the system and that of the standard match, O denotes the condition when the boundary of the system is outside that of the standard and I denotes the condition when the boundary of the system is inside that of the standard. Therefore, the MI, IM, II, MO, OM, IO, OI and OO error types were used to evaluate error distribution. The relative error rate (RER) and the absolute error rate (AER) were computed in error analysis, the relative error rate was compared with all error types, and the absolute error rate was compared with the standard. In the overall error distribution, the main error type was “*IM*” and “*MI*” and the CRF+CAM can significantly reduce those two error types.

Table 5. Distribution of Error Types on Propositional Term Extraction

Error Type	CRF		CRF+CAM		CRF		CRF+CAM	
	RER	AER	RER	AER	RER	AER	RER	AER
	All Data				Speech Data			
MI	24.62	6.11	18.00	2.90	24.90	6.41	20.30	3.03
IM	36.48	8.72	28.50	4.88	38.22	8.06	32.50	4.08
II	18.67	4.96	23.40	3.88	12.37	2.88	14.80	2.05
MO, OM, IO, OI	7.49	3.08	12.50	1.07	10.50	2.46	12.85	1.85
OO	12.74	2.91	17.60	2.08	14.01	4.55	19.55	2.53
	Language Data				Multimedia Data			
MI	23.11	4.03	18.50	2.67	19.18	6.58	17.25	4.64
IM	31.25	9.08	28.50	3.56	25.72	9.00	19.10	4.05
II	26.48	7.50	31.00	4.07	36.34	10.63	34.34	8.30
MO,OM,IO,OI	8.12	1.03	12.45	1.89	6.42	5.00	10.09	1.53
OO	11.04	2.06	9.55	1.20	12.34	4.85	19.22	3.85

4. Conclusion

This study has presented a conditional random field model and a combined association measure approach to propositional term extraction from research abstracts. Unlike previous approaches using POS patterns and statistics to extract NP-based multiword terms, this research considers lexical cohesion and context information, integrating CRFs and CAM to extract single or multiword propositional terms. Experiments demonstrated that in each corpus, both CRF inside and outside tests showed an improved performance for imperfect hits. The proposed approach further effectively enhanced the propositional term boundaries by the combined association measure approach which calculates the cohesiveness of words. The conditional random field model initially detects propositional terms based on their local and global features, which includes statistical, syntactic, semantic, morphological, and structural level information. Experimental results also showed that different multi-level features played a key role in CRF propositional term detection model for different domain data.

References

[1] U. M. Connor, *Contrastive Rhetoric: Cross-Cultural Aspects of Second Language Writing* U.K.: Cambridge Applied Linguistics, 1996.

[2] C. Jacquemin and D. Bourigault, "Term Extraction and Automatic Indexing," in *Oxford Handbook of Computational Linguistics*, M. Ruslan, Ed. Oxford: Oxford University Press, 2003, pp. 599-615.

[3] C.-K. Cheng, *How to Write a Scientific Paper?* Taipei: Hwa Kong Press, 2003.

[4] S. Ananiadou, "A Methodology for Automatic Term Recognition," in *15th Conference on Computational Linguistics - Volume 2*, Kyoto, Japan, 1994, pp. 1034-1038.

[5] F. J. Damerau, "Generating and Evaluating Domain-Oriented Multi-word Terms From Texts," *Inf. Process. Manage.*, vol. 29, pp. 433-447, 1993.

[6] C. Enguehard and L. Pantera, "Automatic Natural Acquisition of a Terminology," *Journal of*

Quantitative Linguistics, vol. 2, pp. 27-32, 1995.

[7] K. T. Frantzi, S. Ananiadou, and H. Mima, "Automatic Recognition of Multi-word Terms: the C-value/NC-Value Method," *Int. J. on Digital Libraries*, vol. 3, pp. 115-130, 2000.

[8] G. Nenadić, S. Ananiadou, and J. McNaught, "Enhancing Automatic Term Recognition through Recognition of Variation," in *20th international conference on Computational Linguistics* Geneva, Switzerland: Association for Computational Linguistics, 2004.

[9] L. A. Ramshaw and M. P. Marcus, "Text Chunking Using Transformation-Based Learning," in *Third Workshop on Very Large Corpora*, 1995, pp. 82-94.

[10] J. Lafferty, A. Mccallum, and F. Pereira, "Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data," in *ICML '01: Proceedings of the Eighteenth International Conference on Machine Learning*, 2001, pp. 282-289.

[11] G. Dias, "Multiword Unit Hybrid Extraction," in *ACL 2003 Workshop on Multiword Expressions: Analysis, Acquisition and Treatment - Volume 18*, 2003, pp. 41-48.

[12] Association for Computing Machinery, Inc., *The ACM Computing Classification System [1998 Version],* New York: ACM. Available: http://www.acm.org/class/1998/. [Accessed: June 17, 2006]

[13] I. Niles and A. Pease, *Suggested Upper Merged Ontology (SUMO) Mapping to WordNet*, Piscataway NJ: IEEE. Available: http://sigmakee.cvs.sourceforge.net/sigmakee/KBs/WordNetMappings/. [Accessed: 2004]

[14] The School of Linguistics and Applied Language Studies at Victoria University of Wellington, *Academic Words List*, Wellington: Victoria University of Wellington. Available: http://www.vuw.ac.nz/lals/research/awl/. [Accessed: June 17, 2006]

[15] E. Charniak, *Eugene Charniak's Parser*, Providence: Brown University. Available: http://cs.brown.edu/~ec/. [Accessed: June 1, 2006]

[16] J. Nocedal, "Updating quasi-Newton Matrices with Limited Storage," *Mathematics of Computation*, vol. 35, pp. 773-782, 1980.

[17] F. Sha and F. Pereira, "Shallow Parsing with Conditional Random Fields," in *2003 Human Language Technology Conference and North American Chapter of the Association for Computational Linguistics (HLT/NAACL-03)*, Edmonton, Canada, 2003, pp. 213-220.

[18] S. C. Lee, "Probabilistic Segmentation for Segment-Based Speech Recognition." M. S. thesis, Massachusetts Institute of Technology, MA, U.S.A., 1998.

[19] G. Nenadić, I. Spasić, and S. Ananiadou, "Automatic Acronym Acquisition and Term Variation Management within Domain-specific Texts," in *Third International Conference on Language Resources and Evaluation (LREC2002)*, Las Palmas, Canary Islands, Spain, 2002, pp. 2155-2162.

[20] S. Jones and J. Sinclair, "English Lexical Collocations: A Study in Computational Linguistics," *Cahiers de Lexicologie*, vol. 23, pp. 15-61, 1974.

[21] G. Dias, "Extraction Automatique d'Associations Lexicales àpartir de Corpora." Ph. D dissertation, DI/FCT New University of Lisbon, Lisbon, Portugal, and LIFO University, Orléans, France , 2002.

[22] K. Taku, *CRF++: Yet Another CRF toolkit 0.44*. Available: http://crfpp.sourceforge.net/. [Accessed: Oct 1, 2006]

[23] A. Lauriston, "Criteria for Measuring Term Recognition," in *Seventh Conference on European Chapter of the Association for Computational Linguistics*, Dublin, Ireland, 1995, pp. 17-22.

[24] K.-M. Park, S.-H. Kim, H.-C. Rim, and Y.-S. Hwang, "ME-based Biomedical Named Entity Recognition Using Lexical Knowledge," *ACM Transactions on Asian Language Information Processing (TALIP)*, vol. 5, pp. 4-21, 2006.