Identifying Coordinated Compound Words for Vietnamese Word Segmentation

Ngoc Anh, Tran
Dept. Information Technology
Le Quy Don Technical University
Hanoi, Vietnam
anhtn69@gmail.com

Thanh Tinh, Dao
Dept. Information Technology
Le Quy Don Technical University
Hanoi, Vietnam
tinhdt@mta.edu.vn

Phuong Thai, Nguyen
Dept. Information Technology
UET, Vietnam National University
Hanoi, Vietnam
thainp@vnu.edu.vn

Abstract— **This paper proposes a dictionary-based method for determining coordinated compound words in Vietnamese. The main idea to determine whether two contiguous simple words in a text forms a coordinated compound word is based on their properties, part-of-speeches and the similarity between their definitions in the dictionary of the Vietnamese Computational Lexicon (VCL). We also based on the sets of synonym and antonym to identify, recognize, and establish a list of coordinated compound words (coordinated di-syllable phrases). We have used a number of rules to identify 3 or 4 syllable phrases/idioms based on relations of coordinated di-syllable phrases. We carried out two major experiments: one for identifying and creating a list of coordinated compounds, the other for improving the accuracy of Vietnamese word segmentation. The second experiment showed that the word segmentation F-scores increases from 0.11% to 0.41% (the error rate decreases from 3.32% to 12.6%). This is a new approach and highly practical value.**

Keywords: new word; coordinated compound words; Vietnamese Computational Lexicon; word segmentation; similarity;

I. INTRODUCTION

The determination of word boundaries is considered the first step in most applications of natural language processing. In Vietnamese, a word is often composed of one or more syllables, so the space does not distinguish the words like English and other languages. On the other hand, the word boundary identification depends on context, for example, the left and the right words. Moreover, the task also depends on the structural properties of Vietnamese complex words such as reduplicative, subordinated, coordinated and non-related. Thus, the determination of Vietnamese word boundaries is a challenge.

Generally, Vietnamese words are classified as Figure 1:

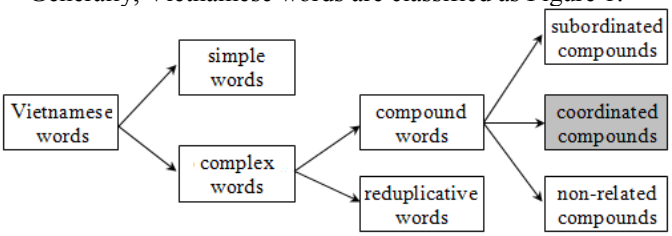

Figure 1. Classification of Vietnamese words

Vietnamese compound words are composed of two or more syllables. Their syllables (constituents) are often related with each other grammatically and semantically. There are three main types of compound words: coordinate compounds, subordinate compounds and non-related compounds regarding to grammatical and semantic relationships. In this paper, we only focus on a special form of Vietnamese compound words, the coordinated compound.

a. Grammatical relationship in coordinated compounds

In coordinate compounds, each constituent is equal in terms of grammatical relationship and the constituents often belong to the same part-of-speech class. However, their part of speech is not fixed. It can vary according to their locations in the sentence. For instance, "chính phụ" may be a noun or adjective. E.g.

1. "*Chính phụ* là một từ ghép đẳng lập." (noun)
2. "Đẳng lập là một từ ghép *chính phụ*." (adjective)

b. Semantic relationship in coordinated compounds

Each word with a separate meaning goes together, referring to a unique entity such as "đất nước" denotes country and "mưa gió" denotes the changes of weather. The meanings of both elements are "combined" or "synthetized" to form the meaning of the compound. However, there are some cases that the compound word carries the meaning of one element only as "nhà cửa" denotes houses in general not doors. Moreover, they are sometimes very idiomatic when separate meaning of each element does not contribute into the meaning of the compound, such as "đất nước" (country) or "gan dạ" (brave).

In our approach, we only consider coordinated compounds whose elements are synonym, or antonym, or similar. Elements are simple words in the VCL (Vietnamese Computational Lexicon) dictionary. The VCL dictionary has 41,734 words and meanings (31,158 words) that contains important information about words such as: morphological information (word type: simple, compound), syntactic information (POS - part-of-speech, sub-POS), and semantic information (categorial meaning, synonym, antonym, definition, example)

Some examples as follows:

1. "xinh" (pretty) and "đẹp" (pretty) are two simple words; for grammatical, they are the same adjectives; for semantic, they are synonyms. Hence, "xinh đẹp" (pretty) is a coordinate compound.
2. "tiến" (advance) and "thoái" (detreat) are two simple words; for grammatical, they are the same verbs; for semantic,

they are antonyms. Hence, "tiến thoái" (advance and detreat) is a coordinate compound.

3. "cha" (father) and "mẹ" (mother) are two simple words; for grammatical, they are the same nouns; for semantic, they are near meaning and their definitions as follows:

+ The definition of "cha" is "người đàn ông có con, trong quan hệ với con" (the man had children, in relation to his child)

+ The definition of "mẹ" is "người phụ nữ có con, trong quan hệ với con" (the woman had children, in relation to her child)

Clearly, their definitions are similar. Hence, "cha mẹ" (parent) is a coordinate compound.

By the above approach, we can identify two contiguous simple words which compose a coordinated compound word. And this also is one of determining ways of new words in Vietnamese.

The rest of this paper is structured as follows: in Section II, we show properties of coordinated compounds. Section III presents how to measure the similarity of two simple words using their definitions. Section IV, rules to identify 3 or 4-syllables phrases. In Section V, experimental results and evaluation. Finally, conclusions are presented in Section VI.

II. PROPERTIES OF COORDINATED COMPOUNDS IN VIETNAMESE

In this section, we show how to verify two contiguous simple words (s_A s_B) is a coordinated compound word (CC). If two simple words (s_A s_B) have coordinated properties, we call they are a coordinated di-syllable phrase (CD). And if the CD is used in the fact text then CD is a CC. We consider some properties as follows:

A. Property of grammatical relationship

They are the same grammatical relationship: POS (part-of-speech) and sub-POS (syntax information) including: noun (N), verb (V), adjective (A), numeral (M), pronoun (P) and prepositions (E).

We have some rules, such as:

(1) N = N + N : "nhà cửa", "gà vịt", "núi đồi",...

* "nhà"(house) + "cửa"(door) = "nhà cửa" (house)
* "gà"(chiken) + "vịt"(duck) = "gà vịt" (poultry)
* "núi"(mountain) + "đồi"(hill) = "núi đồi"(mountain & hill)

(2) V = V + V : "chạy nhảy", "nạo vét", "đấm đá",...

* "chạy" (run) + "nhảy" (jump) = "chạy nhảy" (run & jump)
* "nạo" (scrape) + "vét" (dredge) = "nạo vét" (dredge)
* "đấm" (punch) + "đá" (kick) = "đấm đá" (punch & kick)

(3) A = A + A : "già trẻ", "nhỏ bé", "xinh đẹp",...

* "già" (old) + trẻ (young) = "già trẻ" (young and old)
* "nhỏ" (small) + "bé" (tiny) = "nhỏ bé" (small)
* "xinh"(pretty) + "đẹp"(beautiful) = "xinh đẹp"(beautiful)

(4) M = M + M : "một vài", "đôi ba", "ba bảy",...

* "một" (one) + "vài" (some) = "một vài" (some)
* "đôi" (double) + "ba" (triple) = "đôi ba" (a few)
* "ba" (three) + bảy (seven) = "ba bảy" (more than one)

(5) P = P + P : "đó đây", "này nọ", "kia kìa",...

* "đó" (there) + "đây" (here) = "đó đây" (here and there)
* "này" (this) + "nọ" (that) = "này nọ" (this and that)
* "kia" (there) + "kìa" (there) = "kia kìa" (there)

(6) E = E + E : "trong ngoài", "trên dưới", "trước sau",...

* "trong" (inside) + "ngoài" (outside) = "trong ngoài" (inside & outside)
* "trên" (above) + "dưới" (below) = "trên dưới" (above & below)
* "trước" (before) + "sau" (after): "trước sau" (before & after)

B. Property of semantic relationship

The meanings of both elements are "combined" or "synthetized" to form the meaning of the compound with semantic relationships as follows:

(1) Synonym: "chờ đợi", "khoẻ mạnh", "xinh đẹp",...

* Synonym of "đợi" (wait) is "chờ" (wait): "chờ đợi" (wait)
* Synonym of "mạnh" (strong) is "khoẻ" (healthy): "khoẻ mạnh" (healthy)
* Synonym of "đẹp" (beautiful) is "xinh" (pretty): "xinh đẹp" (beautiful)

(2) Antonym: giàu nghèo, tốt xấu, cha mẹ,...

* Antonym of "giàu" (rich) is "nghèo" (poor): "giàu nghèo" (rich and poor)
* Antonym of "tốt" (good) is "xấu" (bad): "tốt xấu" (good and bad)
* Antonym of "cha" (father) is "mẹ" (mother): "cha mẹ" (parent)

(3) Similar: ăn uống, sông suối, núi đồi,... We use their definitions to compare

* Definition of "ao" (pond) and "hồ" (lake):

+ Def. of "ao" (pond) is "chỗ đào sâu xuống đất, thường ở gần nhà, để giữ nước nuôi cá, thả bèo, trồng rau, v.v" (digged place on the ground, usually near the house, to keep the fish water, drop dirt, vegetable, etc.)

+ Def. of "hồ" (lake) is "nơi đất trũng chứa nước, thường là nước ngọt, tương đối rộng và sâu, nằm trong đất liền" (lowlands where hold fresh water, relatively wide and deep inland)

* Definition of "sông" (river) and "suối" (stream):

+ Def. of "sông" (river) is "dòng nước tự nhiên tương đối lớn, chảy thường xuyên trên mặt đất, thuyền bè thường đi lại được" (the large natural flow of water, frequent running on the ground, usually ships move on it)

+ Def. of "suối" (stream) is "dòng nước tự nhiên ở miền đồi núi, chảy thường xuyên hoặc theo mùa, do nước mưa hoặc nước ngầm chảy ra ngoài mặt đất tạo nên" (the natural flow of water in the areas of mountains and hills is made by rainwater or goundwater flow on or under the ground)

C. Property of permutation

If (s_A s_B) is a CC then (s_B s_A) is a CD.

If (s_A s_B) is a CD then (s_B s_A) is also a CD.

* "chờ đợi" (wait) is a CC, so "đợi chờ" (wait) is a CD.
* "áo quần" (clothes) is a CC, so "quần áo" is a CD.
* "cha mẹ" (parent) is a CD, so "mẹ cha" is a CD.

D. Property of transitive

If (s_A s_B) and (s_B s_C) are CCs then (s_A s_C) is a CD.

If (s_A s_B) and (s_B s_C) are CDs then (s_A s_C) is a CD.

* "mong chờ" (expect) and "chờ đợi" (wait) are CCs, so "mong đợi" (expect) is a CD.

* "xinh tươi" (pretty) and "xinh đẹp" (beautiful) are CCs, so "tươi đẹp" (beautiful) is a CD.

* "tuyển chọn" (choose) and "lựa chọn" (choose) are CDs, so "tuyển lựa" (choose) is a CD.

E. Property of existence

By filtering Vietnamese sentences of online / offline raw-corpus based on syllables (online on the internet) that contain two coordinated syllables, we can determine their existence. In the fact, they often appear side by side with some frequency and we can use the mutual information of coordinated di-syllable phrases (MI). Geometrically, we can describe the following schema: we call $I\text{sw}_\text{A}$ and $I\text{sw}_\text{B}$ are information about syllables sw_A and sw_B in the corpus.

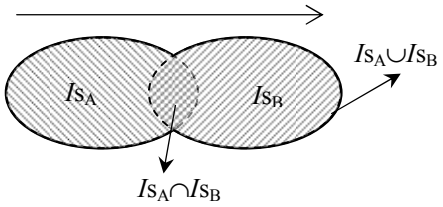

Figure 2. Schema of mutual information of two syllables (s_A s_B)

So, the mutual information of syllables can be defined as the measure the linking between two syllables:

$$MI(s_A s_B) = \frac{|\, Is_A \cap Is_B \,|}{|\, Is_A \cup Is_B \,|} \tag{1}$$

Here, we have:

$|\, Is_A \,| = C(s_A)$ and $|\, Is_B \,| = C(s_B)$

$|\, Is_A \cap Is_B \,| = C(s_A\, s_B)$

$|\, Is_A \cup Is_B \,| = |\, Is_A \,| + |\, Is_B \,| - |\, Is_A \cap Is_B \,|$

$= C(s_A) + C(s_B) - C(s_A\, s_B)$

Thus,

$$MI(s_A\, s_B) = \frac{C(s_A\, s_B)}{C(s_A) + C(s_B) - C(s_A\, s_B)} \tag{2}$$

where,

$MI(s_A\, s_B)$ is linking of two syllables $(s_A\, s_B)$

$C(s_A\, s_B)$ is the count of syllable bigram $(s_A\, s_B)$

$C(s)$ is the count of syllable unigram (s).

If $(s_A\, s_B)$ is a CD and $MI(s_A\, s_B)$ is greater than threshold MI_0 then $(s_A\, s_B)$ is a CC.

All above properties, properties (A) and (E) are used as conditions to determine results in properties (B), (C) and (D).

III. SIMILARITY MEASURE OF TWO SIMPLE WORDS

We know that, part II.B (property of semantic relationship) shows a form of semantic relationship (similar) need to be computed by the similarity between definitions of two simple words. However, until now, a Vietnamese wordnet is no available yet. So, we compute the lower level of semantics, syllable level.

When the word boundaries aren't separated, Vietnamese text is a sequence of syllables. So, similarity measure of two texts (paragraphs or sentences or phrases) based on syllables.

A. Similarity based on approximate string matching by distance

To measure the similarity, we use algorithms of approximate string matching by distance with syllables.

Given two X and Y texts are two sequences of syllables

$X = \{x_1, x_2, x_3, ..., x_\text{n}\}$ and $Y = \{y_1, y_2, y_3, ..., y_\text{m}\}$;

where: $|\, X \,| = n$; $|\, Y \,| = m$;

We can apply some formulas of distance between two X and Y texts with syllables to compute their similarity.

1) Similarity based on Edit Distance (ED)[15]

$$Sim_{ED}(X,Y) = \frac{\max(|\, X \,|, |\, Y \,|) - ED(X,Y)}{\max(|\, X \,|, |\, Y \,|)} \tag{3}$$

E.g. x = "con chó cắn con mèo" (the dog bites the cat)

y = "con mèo cắn con chuột" (the cat bites the mouse)

By order, we have two sequences of syllables as follows:

X = (con, chó, cắn, con, mèo) and Y = (con, mèo, cắn, con, chuột)

Executing algorithm of edit distance[15] with X and Y.

We have: $ED(X,Y) = 2$, hence:

$$Sim_{ED}(X,Y) = \frac{5-2}{5} = 0.6$$

2) Similarity based on Longest Common Subsequence LCS

$$Sim_{LCS}(X,Y) = \frac{2\,|\, LCS(X,Y) \,|}{|\, X \,| + |\, Y \,|} \tag{4}$$

For X = (con, chó, cắn, con, mèo) and Y = (con, mèo, cắn, con, chuột)

We have: $LCS(X,Y)$ = (con, cắn, con), so $|LCS(X,Y)| = 3$.

Hence: $$Sim_{LCS}(X,Y) = \frac{2\times 3}{5+5} = 0.6$$

Similarly, we can apply similarity to other distances [15] as: *Hamming Distance* (*HD*), *Damerau-Levenshtein Distance* (*DLD*). In addition, we can use the measure of *Jaro-Winkler Distance* (*JW*), *Trigram Distance* (*Trigram*) or *Ratcliff / Obershelp* for computing the similarity between two texts.

B. Similarity based on co-occurrence measures

Given two X and Y texts are two sequences of syllables

$X = \{x_1, x_2, x_3, ..., x_\text{n}\}$ and $Y = \{y_1, y_2, y_3, ..., y_\text{m}\}$;

where: $|\, X \,| = n$; $|\, Y \,| = m$;

$|\, X \cap Y \,|$ = number of syllables co-occur in X and Y

$|\, X \cup Y \,| = |\, X \,| + |\, Y \,| - |\, X \cap Y \,|$

a. Dice similarity:

$$Sim_{Dice}(X,Y) = 2\frac{|\, X \cap Y \,|}{|\, X \,| + |\, Y \,|} \tag{5}$$

b. Jaccard similarity:

$$Sim_{Jaccard}(X,Y) = \frac{|\, X \cap Y \,|}{|\, X \cup Y \,|} \tag{6}$$

For X = (con, chó, cắn, con, mèo) and Y = (con, mèo, cắn, con, chuột)

We have: $|\, X \,| = 5$; $|\, Y \,| = 5$; $|\, X \cap Y \,| = 4$; $|\, X \cup Y \,| = 6$.

Hence,

$$S_{Dice}(X,Y) = 2\times\frac{|\, X \cap Y \,|}{|\, X \,| + |\, Y \,|} = 2\times\frac{4}{10} = 0.8$$

$$S_{Jaccard}(X,Y) = \frac{|\, X \cap Y \,|}{|\, X \cup Y \,|} = \frac{4}{6} \approx 0.67$$

C. Similarity based on Vector Space Model (VSM)

Texts are normalized by the length as a form of vectors:

$$\begin{cases} \vec{x} = (x_1, x_2, ..., x_m) \\ \vec{y} = (y_1, y_2, ..., y_m) \end{cases}$$

We use cosine similarity (vector of frequency):

$$Sim_{Cos}(X,Y) = \cos(\vec{x},\vec{y}) = \frac{\vec{x}.\vec{y}}{|\vec{x}|.|\vec{y}|} = \frac{\sum_{k=1}^{m} x_k y_k}{\sqrt{\sum_{k=1}^{m} x_k^2}\sqrt{\sum_{k=1}^{m} y_k^2}} \quad (7)$$

x = "con chó cắn con mèo" and y = "con mèo cắn con chuột"

Cx = (con, chó, cắn, mèo) and *Cy* = (con, mèo, cắn, chuột).

With *Cxy* = *Cx* ∪ *Cy* = (con, chó, cắn, mèo, chuột) is sequence of common syllables.

After sorted *Cxy*, *Cxy* = (cắn, chó, chuột, con, mèo)

By frequency of elements:

$$\begin{cases} X = (\text{cắn:1, chó:1, chuột:0, con:2, mèo:1}) \\ Y = (\text{cắn:1, chó:0, chuột:1, con:2, mèo:1}) \end{cases}$$

They are equivalent to:

$$\begin{cases} \vec{x} = (1, 1, 0, 2, 1) \\ \vec{y} = (1, 0, 1, 2, 1) \end{cases}$$

Hence,

$$Sim_{Cos}(X,Y) = \frac{1+0+0+4+1}{\sqrt{(1+1+0+4+1)}\sqrt{(1+0+1+4+1)}} = \frac{6}{7} \approx 0.86$$

IV. Identifying 3 or 4-syllables phrases

A. Identifying 3-syllables phrase with coordinated property

1) For 3-syllables phrase with coordinated tri-syllables

In fact text, if exist a sequence of 3 syllables (s_A s_B s_C), they coordinate with each others, eg. (s_A s_B) and (s_B s_C) are CDs, then they are a coordinated tri-syllables phrase.

For example, with a sequence "*anh chị em*" is used in fact text, we have "*anh chị*" and "*chị em*" are CCs, so "*anh chị em*" is a coordinated tri-syllables phrase.

2) For 3-syllables phrase with coordinated di-syllable

In many cases, if exist a sequence of 3 syllables (s_A s_B s_C), where, (s_A s_C) and (s_B s_C) are two subordinated compounds, (s_A s_B) is a CC, then (s_A s_B s_C) is a 3-syllables phrase.

A given sequence, "*thầy cô giáo*" is used in fact text, where, "*thầy giáo*" and "*cô giáo*" are two subordinated compounds, and "thầy cô" is a coordinated compound, so "*thầy cô giáo*" is a 3-syllables phrase based on coordinated di-syllable.

B. Identifying 4-syllables phrase with coordinated property

1) For 4-syllables phrase with coordinated di-syllable

In fact text, if exist a sequence of 4 syllables (s_A s_B s_A s_C), where, (s_B s_C) is a coordinated di-syllable phrase, then (s_A s_B s_A s_C) is a 4-syllables phrase/idiom with the coordinated di-syllable.

For example, a sequence "*hết lòng hết dạ*" is used in a text, where, "*lòng dạ*" is a coordinated di-syllable phrase. Hence, the sequence "*hết lòng hết dạ*" is a 4-syllables phrase/idiom with a coordinated di-syllable phrase.

2) For 4-syllables phrase with 2 coordinated di-syllables

In fact, if exist a sequence of 4 syllables (s_A s_C s_B s_D), where, (s_A s_B) and (s_C s_D) are two coordinated di-syllable phrases, then (s_A s_C s_B s_D) is a 4-syllables phrase/idiom with two alternate coordinated di-syllable phrases.

For example, a given sequence "*cơm no áo ấm*" is used in a text, where, "*cơm áo*" and "no ấm" is two coordinated di-syllable phrases. So, "*cơm no áo ấm*" is a 4-syllables phrase / idiom with two alternate coordinated di-syllable phrases.

V. Experiments

A. The corpus and dictionary for testing and assessment

1) The dictionaries

- VCL dictionary [12]: this dictionary has 41,734 items and meanings with 31,158 words, that contains all the information of the morphological (word type: simple word, compound), the syntax (POS - part-of-speech, sub-POS), the semantic (categorial meaning, synonym, antonym, definition, example).

2) The corpus for training and testing

SP73 Corpus [13]: the VietTreeBank Corpus includes 70,000 sentences, with a total of 1,547,387 segmented words, the size of 10MB (SP7.3, the project of KC.01.01/06-10). We split this corpus to 2 parts:

- Training corpus (70% SP73 Corpus): the training corpus includes about 49,000 sentences with a total of 1,094,099 words, the size 7.2MB. This corpus is used for the ngram statistics, computing probabilities of word bigram and the mutual information of syllables. The statistical and training results are saved into a knowledge database file for later use.

- Testing corpus (30% SP73 Corpus): the testing corpus includes about 21,000 sentences with a total of 453,288 words, the size 2.9MB.

Before using, we take spelling checker for SP73. We have checked, edited and merged almost coordinated compound words for SP73 corpus.

3) The assessment of accuracy

The assessment of accuracy based on the below way:

+ Recall (R): number of correctly segmented words divided by the total of words in the corpus.

+ Error Recall (ErrR): number of errors by recall

+ Precision (P): number of correctly segmented words divided by the total of segmented words in the solution.

+ Balance F-score (F):

$$F = \frac{2PR}{P+R} \quad (8)$$

B. Experiment with identifying coordinated compounds

1) Searching pairs of synonyms

TABLE I. Some pairs of synonyms (simple words)

N.	Pairs of synonyms	Pos	N.	Pairs of synonyms	Pos
1.	ẵm bồng	V	11.	canh gác	V
2.	ầm rầm	A	12.	câu cú	N
3.	ẩn nấp	V	13.	chiếm đoạt	V
4.	ba bố	N	14.	dơ bẩn	A
5.	bại liệt	V	15.	đảng phái	N
6.	bằng phẳng	A	16.	đấu nối	V
7.	bắt buộc	V	17.	đói khát	V
8.	bè phái	N	18.	gan lì	A
9.	cạm bẫy	N	19.	gặt hái	V
10.	cầm cố	V	20.	hình ảnh	N

Searching a list of pairs of synonyms (simple words) from VCL dictionary, it includes 3446 pairs of simple words.

2) Searching pairs of antonyms

Searching a list of pairs of antonyms (simple words) from VCL dictionary, it includes 550 pairs of simple words.

TABLE II. SOME PAIRS OF ANTONYMS (SIMPLE WORDS)

N.	Pairs of antonyms	Pos	N.	Pairs of antonyms	Pos
1.	ác thiện	A	11.	cậu mợ	N
2.	âm dương	A	12.	cha mẹ	N
3.	âm dương	N	13.	chắc lép	A
4.	ẩm khô	A	14.	chẵn lẻ	A
5.	ân oán	N	15.	chết sống	V
6.	anh chị	N	16.	chìm nổi	V
7.	ba mẹ	N	17.	chính phụ	A
8.	bật tắt	V	18.	co giãn	V
9.	béo gầy	A	19.	công tội	N
10.	cao thấp	A	20.	cũ mới	A

3) Searching pairs of the same definition words

Searching a list of pairs of simple words which the same definition from VCL, it includes 482 pairs of simple words.

TABLE III. SOME PAIRS OF THE SAME DEFINITION WORDS

N.	Pairs of the same def.	Pos	N.	Pairs of the same def.	Pos
1.	ba tía	N	11.	chít chịt	V
2.	bằm vằm	V	12.	coi ngó	V
3.	bắp bẹ	N	13.	cọp hùm	N
4.	bệnh bịnh	N	14.	dày dầy	A
5.	béo bẹo	V	15.	dấp nhấp	V
6.	béo nhéo	V	16.	đậu đỗ	V
7.	bóng hình	N	17.	đó đây	P
8.	bừa phứa	A	18.	găm ghim	V
9.	cài gài	V	19.	kêu van	V
10.	chặt chém	V	20.	lãi lời	N

4) Experiment with some similarity measures

TABLE IV. SIMILARITY OF SOME PAIRS OF SIMPLE WORDS

N.	Pairs	Pos	SimED	SimLcs	SimDic	SimJac	SimCos
1.	đông tan	V	0.7778	0.7778	1.0000	1.0000	1.0000
2.	trên dưới	E	0.9048	0.9048	1.0000	1.0000	1.0000
3.	nam nữ	N	0.8571	0.8571	1.0000	1.0000	1.0000
4.	sấp ngửa	A	0.4737	0.6316	1.0000	1.0000	1.0000
5.	anh em	N	0.7037	0.7308	0.9615	0.9259	0.9866
6.	bắc nam	N	0.8696	0.8696	0.9565	0.9167	0.9726
7.	cứng rắn	A	0.9412	0.9697	0.9697	0.9412	0.9701
8.	hữu tả	N	0.6735	0.7579	0.9474	0.9000	0.9687
9.	vợ chồng	N	0.7619	0.7619	0.9524	0.9091	0.9655
10.	đúng sai	A	0.9167	0.9565	0.9565	0.9167	0.9636
11.	trong ngoài	N	0.7407	0.7925	0.9434	0.8929	0.9542
12.	dần thìn	N	0.9184	0.9184	0.9388	0.8846	0.9538
13.	hở kín	A	0.9091	0.9524	0.9524	0.9091	0.9535
14.	ông bà	N	0.9500	0.9500	0.9500	0.9048	0.9500
15.	hỏi ngã	N	0.9412	0.9412	0.9412	0.8889	0.9474

Experiment and considering a list of some simple words to evaluate similarity for pairs of simple words.

5) Experiment with existent of coordinated di-syllables

We use mutual information of coordinated di-syllable phrase based on Google search engine to determine AB (calculating with 20 first search results) is a coordinated compound?

With the above experiment results (1,2,3,4), we have filtered and removed the coordinated di-syllable phrases which do not exist in fact. (See Table V).

TABLE V. SOME COORDINATED DI-SYLLABLE PHRASES

N.	Pairs	MI	N.	Pairs	MI	N.	Pairs	MI
1.	ẩn nấp	0.8235	11.	ác thiện	0.0000	21.	đông tan	0.0000
2.	nấp ẩn	0.0000	12.	thiện ác	0.3404	22.	tan đông	0.0000
3.	bè phái	0.8125	13.	âm dương	0.6176	23.	trên dưới	0.2889
4.	phái bè	0.0000	14.	dương âm	0.0000	24.	dưới trên	0.0000
5.	cạm bẫy	0.6774	15.	ẩm khô	0.0000	25.	nam nữ	0.2131
6.	bẫy cạm	0.0000	16.	khô ẩm	0.0000	26.	nữ nam	0.0267
7.	bóng hình	0.3390	17.	ân oán	0.8276	27.	bắc nam	0.1636
8.	hình bóng	0.4211	18.	oán ân	0.0000	28.	nam bắc	0.0926
9.	bừa phứa	0.8636	19.	cao thấp	0.2941	29.	vợ chồng	0.4833
10.	phứa bừa	0.0000	20.	thấp cao	0.0000	30.	chồng vợ	0.0345

TABLE I & III TABLE II TABLE IV

As you see, MI(ẩn nấp) = 0.8235, that means "ẩn nấp" is a CC, but MI(nấp ẩn) = 0.0000, "nấp ẩn" does not exist. Some cases with similarity of CDs is very high, for example, "đông tan", all SimCos, SimDice, SimJacard equal to 1 (Table IV), but MI("đông tan") or MI("tan đông") equal to 0 (Table V), that means they do not exist.

Finally, we have collected a list of about 4500 coordinated di-syllable phrases (before, [14] shows 3979 CDs phrases)

C. Experiment for identifying 3- and 4-syllables phrases

Based on list of coordinated di-syllable phrases collected, we have identified 3- and 4-syllables phrases on VietTreeBank Corpus. And, we detected many 3- and 4-syllables phrases based on coordinated properties.

After that, we have checked and modified many errors for VietTreeBank Corpus (coordinated di-syllable phrases, errors of 3- or 4-syllables phrases).

In Table VI, we describe some 3- and 4-syllables phrases

TABLE VI. SOME 3- OR 4-SYLLABLES PHRASES

N.	3-syllables phrases	Pos	N.	4-syllables phrases	Pos
1.	anh chị em	N	11.	bữa rau bữa cháo	N
2.	ca múa nhạc	N	12.	chạy ngược chạy xuôi	V
3.	công nông binh	N	13.	cứu khổ cứu nạn	V
4.	thanh thiếu nhi	N	14.	đồng ra đồng vào	N
5.	cơ xương khớp	N	15.	một mất một còn	A
6.	động thực vật	N	16.	ăn đói mặc rách	V
7.	nạo phá thai	V	17.	chia ba xẻ bảy	V
8.	thanh thiếu niên	N	18.	đi nắng về mưa	V
9.	thầy cô giáo	N	19.	nước mất nhà tan	N
10.	ưu nhược điểm	N	20.	trời cao đất dày	N

D. Experiment for word segmentation

In [11, 16], we had performed some experiment for Vietnamese word segmentation (VWS). In this part, we add module identifying coordinated compound (CC) to improve the accuracy of word segmentation. We call some other moduls for word segmentation: FMM (Forward Maximum Matching), BMM (Backward Maximum Matching), MM (Maximum Matching with balancing), NE (Named Entity recognization), MI (Mutual Information of syllable ngrams), Pb (Probability of word bigrams).

The results in Table VII:

TABLE VII. RESULTS OF WORD SEGMENTATION - VCL 31,158 WORDS

	Integrated methods	ErrR	δEr	R (%)	P (%)	F (%)	δF%
	FMM	20086		95.57	92.11	93.81	
	BMM	19194		95.77	92.30	94.00	
	NE+BMM	8986		98.02	97.17	97.59	
i	NE+MM	*8984*	*961*	*98.02*	*97.17*	*97.59*	
	NE+MM+**CC**	8023	12 %	98.23	97.76	**97.99**	+0.40
ii	NE+MM+MI	*8839*	*990*	*98.05*	*97.20*	*97.62*	
	NE+MM+ **CC**+MI	7849	12.6%	98.27	97.80	**98.03**	+0.41
iii	NE+MM+Pb	*5875*	*189*	*98.70*	*97.99*	*98.34*	
	NE+MM+**CC**+Pb	5686	3.32%	98.75	98.16	**98.45**	+0.11
iv	NE+MM+MI+Pb	*5853*	*189*	*98.71*	*97.99*	*98.35*	
	NE+MM+**CC**+MI+Pb	5664	3.34%	98.75	98.16	**98.46**	+0.11

(i) Only using a dictionary with MM for word segmentation (with CC: the F increases 0.4%, the error rate decreases 12%)

(ii) Using dictionary with MM and a raw-corpus for unsupervised learning of syllable ngrams (with CC: the F increases 0.41%, the error rate decreases 12.6%)

(iii) Using dictionary with MM and segmented word corpus for supervised learning of word bigrams. (with CC: the F increases 0.11%, the error rate decreases 3.32%)

(iv) Using dictionary with MM, mutual information of syllables ngrams and probability of word bigrams.(with CC: the F increases 0.11%, the error rate decreases 3.34%)

So, the errors (by recall) decrease by integrating methods top-down. Clearly, when add module identifying coordinated compound (CC), results of word segmentation are much better than previous. [See two columns δF and δEr on rows (i), (ii), (iii) and (iv)].

VI. CONCLUSIONS

We have presented our research results of determining two contiguous simple words in corpora. Coordinated compound words can be identified by using information in the dictionary of VCL, properties of coordinated compound words, a similarity measure between definitions of simple words in VCL. Our experiments showed that this is an effective way for detecting new words employing Vietnamese features. This method also improves the accuracy of Vietnamese word segmentation.

ACKNOWLEDGEMENT

We would like to express our thanks to Dr. Nguyen Thi Trung Thanh (Institute of Linguistics). She helped us checking and modifying the list of coordinated di-syllable phrases in VCL dictionary and VietTreeBank corpus of 70,000 sentences with segmented words. This paper has been supported by VNU project "Exploiting Very Large Monolingual Corpora for Statistical Machine Translation" (code QG.12.49).

REFERENCES

In Vietnamese:

[1] Ban D. Q., Thung H. V. (2006), *Ngữ pháp tiếng Việt*, Volume 1 & 2, Education P.H., Hanoi.

[2] Can N. T. (1975), *Ngữ pháp tiếng Việt: tiếng, từ ghép, đoản ngữ*, P.H. of University and vocational schools.

[3] Chau D. H. (2004), *Giáo trình Từ vựng học tiếng Việt*, P.H. of University of Education.

[4] Duong N. D. (1971), "Vài nét về những tổ hợp gồm hai yếu tố trái nghĩa trong tiếng Việt", *Ngôn ngữ*, Vol. 2, 1971.

[5] Duong N. D. (1974), "Về các tổ hợp song tiết tiếng Việt", *Ngôn ngữ*, Vol.2, 1974.

[6] Giap N. T. (2010), *Từ vựng học tiếng Việt*, Education P.H., Viet Nam

[7] Hanh H. V. (1984), "Về những nhân tố quy định trật tự các thành tố trong đơn vị song tiết của tiếng Việt", *Ngôn ngữ*, Vol.2, 1984.

[8] Huyen N.T.M., Linh H.T.T., Luong V.X. (2009), "Hướng dẫn nhận diện đơn vị từ trong ngôn ngữ tiếng Việt", Report of SP8.2, Volume 2-VLSP, Project of KC01.01/06-10.

[9] [Phe Hoang], Viet B.K., Thu C.B., Than D., Tue H., Hanh H.V., Chi L.K., Chau N.M., Tram N.N., Nga N.T., Khanh N.T., Khang N.K., Viet P.H., Van T.C., Phuong T.N., Bao V.N., Loc V. (2010), *Từ điển tiếng Việt*, Institute of Linguistics, Vietnam Encyclopedia P.H.

[10] Nghieu V. D. (1999), "Các đơn vị từ vựng song tiết đẳng lập tiếng Việt trong bối cảnh một số ngôn ngữ Đông Nam Á", *Ngôn ngữ*, Vol.5, 1999.

[11] Ngoc Anh Tran, Thanh Tinh Dao, Phuong Thai Nguyen (2011), "Một phương pháp hiệu quả khử nhập nhằng theo ngữ cảnh trong bài toán tách từ tiếng Việt", *Tạp chí Khoa học & Kỹ thuật*, HVKTQS, Vol. 145, 12/2011, pp.50-62

[12] Project KC01.01/06-10 (VLSP), Nhánh đề tài xử lý văn bản (2010), Dictionary of VCL (Vietnamese Computational Lexicon).

[13] Thai N.P., Luong V.X., Huyen N.T.M., Phuong L.H., Thu. D.M., Ngoc N.T.M, Ngan L.K., Van N.M. (2009), Report of SP7.3 - VietTreeBank, Volume 1-VLSP, Project of KC01.01/06-10.

[14] Thanh N. T. T. (2003), *Đặc điểm tổ hợp song tiết đẳng lập tiếng Việt*, Institute of Linguistics, Hanoi, Viet Nam.

In English:

[15] Michel M. Deza and Elena Deza (2009), *Encyclopedia of Distances*, Springer 2009.

[16] Ngoc Anh Tran, Thanh Tinh Dao, Phuong Thai Nguyen (2012), "An effective context-based method for Vietnamese-word segmentation", *First International Workshop on Vietnamese Language and Speech Processing (VLSP2012), In conjunction with 9th IEEE-RIVF Conference on Computing and Communication Technologies*, 2/2012, pp.34-40