TOWARDS A MULTI-OBJECTIVE CORPUS FOR VIETNAMESE LANGUAGE

Vu Hai Quan, Pham Nam Trung, Nguyen Duc Hoang Ha,
Huynh Bao Toan, Le Hoai Bac, Hoang Kiem
Vietnam Natitional Universiy at HCMCT
227 Nguyen Van Cu, HoChiMinh city,Vietnam
vhquan@fit.hcmuns.edu.vn

Abstract

Today, corpus plays an important role in development and evaluation language and speech technologies, such as part of speech tagging, parsing, word sense disambiguation, text categorization, named entity classification, information extraction, question answering, structure discovery (clustering), speech recognition and machine translation systems, etc. One can exploit valuable statistical parameters taken from corpus to train and evaluate those systems. Developing such a corpus has been a challenging work in context that a huge data needed to be processed and annotated. In this paper we first represent our developing method for a multi-objective Vietnamese language corpus, namely **VnCorpus**, together with the description of various kinds of sources from which we have used to build up this database. It then goes on to describe some first experiences in using this corpus for the segmentation of sentences into Vietnamese words and for the recognition of Vietnamese continuous speech. Upon completion the corpus will constitute a valuable resource for research in the fields of computational linguistics, language and speech technologies.

1 Introduction

Vietnamese speech has been created approximately 4000 years, closely related with Indo-European languages. Today there are around 80.000.000 people using this language. The main feature which makes it differ from Western languages is that it belongs among the group of mono - syllable languages. That means it never changes its morphology. In order to express grammatical sense we usually use means of the outside word as grammatical words, order words, etc. The other important feature that makes it differ from Eastern language is that it uses extended Latin based symbols.

In Vietnamese language, the basic is "tiếng". There are totally around 8000 "tiếng" found in Vietnamese modern language [1]. For speech, in the complete form, "tiếng" has following model: (as shown in Fig 1)

Tone			
initial sound	Syllable		
	inter-sound	main sound	final sound

Fig. 1 Structure of "tiếng"

In some cases, "tiếng" can be appeared without final sound, inter-sound, or initial sound. That means main sound and tone are the major components from which "tiếng" is formed. There are 22 initial sounds, 14 main sounds and 10 final sounds coordinating with 6 tones. In order to distinguish "tiếng", initial sound, inter-sound, main sound and final sound are used. In cases that all of them are the same, tones are used. An example of "tiếng" and tones are given in Fig. 2 and Fig. 3.

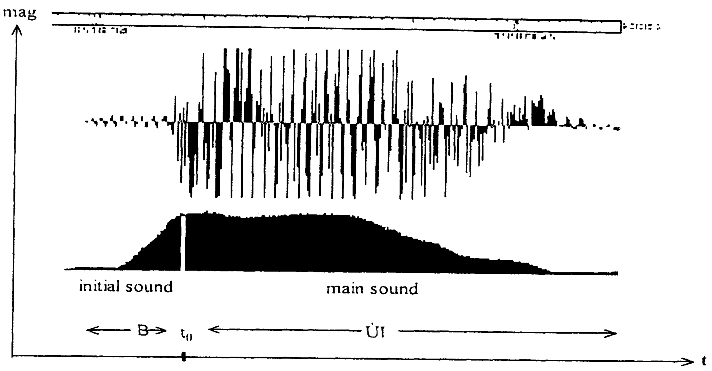

Fig. 2. Structure of Vietnamese "tiếng" "BÙI"

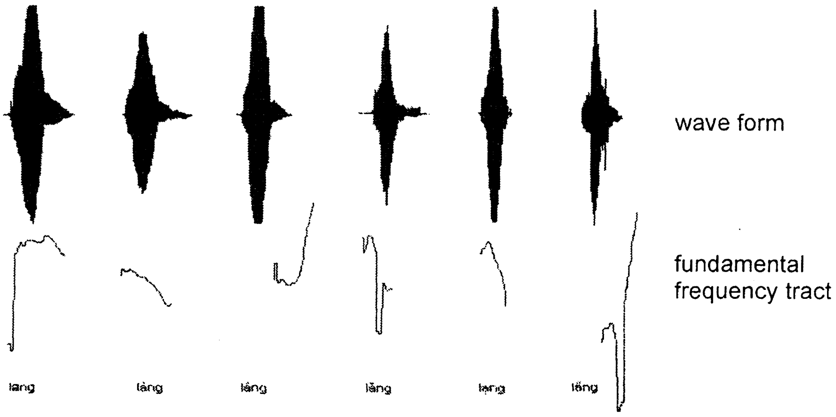

Fig 3. The different tones of Vietnamese

For writing system, "tiếng" usually consists of two main components: consonant (corresponding to initial sound) and syllable (corresponding to inter-sound, main sound and final sound) coordinating with accent in the accent set (corresponding to tones). There are totally 27 consonants (table 1), 434 syllables (a part of them are shown in Table 2) and 6 accents (Table 3).

Table 1. Consonants

b	c	d	đ	g	h	k	l	m	n	p	q	r	s
t	v	x	ph	th	nh	ch	gi	gh	ng	ngh	tr	kh	

Table 3. The accent set

□	\	/	ʔ	~	•

Table 2. Syllables

a oa ac oac ach oach ai oai am oam an oan
ang oang anh oanh ao oao ap oap at oat au
oau ay oay uac uan uang uap uat uay e oe
ec em en oen eng oeng eo oeo ep oep et
oet ue ech uech uên enh uenh uet eu i uy ia
uya ich uych iec iem ien uyn ieng iep iet
uyet ieu yen yeu im in uynh ip uyp it uyt yu
uyu o oc oi om on ong ooc oong op ot uơ ơi
ơm ơn ơp ơt ua uc ui um un ung uoc uoi
uom uon uong uot up ut ưa ưc ưi ưm ưn
ưng ươc ươi ươm ươn ương ươp ươt ươu ưt ưu

In general all Vietnamese words are created from "tiếng". We can have a word with one, two, three, four or even five "tiếng". However not all "tiếng" have meaning. For simply, we can suppose that a single word is a meaning word that contains only one "tiếng". More than 90% of all single words belong to this kind (The others can be determined by hand). Compound word or word sequence is a meaning word which is formed from more than one "tiếng". For example if we make a query to find all compound words that begin from "tiếng" "học" (to study, to learn) from the dictionary we could have at least 79 alternative word sequences in which eight of them are combined from 3 consecutive "tiếng", two of them are created from 4 consecutive "tiếng" and the remainder are formed from two consecutive "tiếng" as shown in Fig 4.

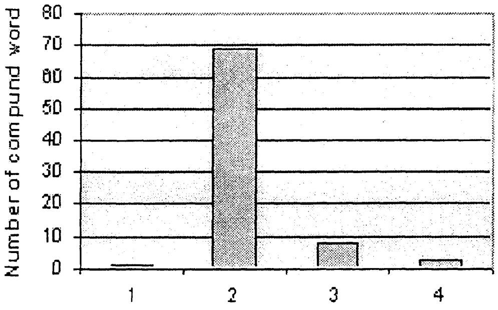

Fig. 4 Compound words from tiếng "học"

The appearance of compound words raises several problems for Vietnamese natural language processing such as the part of speech tagging and parsing because we need a correct segmentation of those compound words in each sentence. Considering the sentence: "Học sinh học sinh vật" (Pupils learn biology) we could have following combination of correct word tags (see Table 4) while the correct segmentation should be "Hoc sinh/*Noun*/ học/*Verb*/sinh vật/*Noun*".

In the previous works, we have introduced some approaches for the recognition of Vietnamese document images [3], handwritten images [4] and also isolated speech words [2]. However we soon recognized that in order to archive a better result, a corpus is needed. In the following sections, we will present the VnCorpus in details and give some first experiences in using this corpus for the segmentation of sentences into correct word sequences and for the recognition of Vietnamese continuous speech.

Table 4. The various of part of speech for a sentence
“Học sinh học sinh vật” - Pupils learn biology

Vietnamese	Tag	English
học	Verb,	to study, to learn
	Proper Name	
học sinh	Noun	pupil
Sinh	Verb	to give birth to sb
sinh học	Noun	biology
sinh vật		
vật	Noun	thing, object
	Verb	to wrestle

2 The VnCorpus design.

From 2002, we began to build the VnCorpus which includes following sources: written texts, spoken corpus and parallel Vietnamese-English texts. The size of the corpus is approximately 100 million words, which is distributed to written texts (80 million single words), spoken (4 million isolated words) and parallel Vietnamese-English texts and spoken (16 million words). The time table of this project is shown in Fig 5.

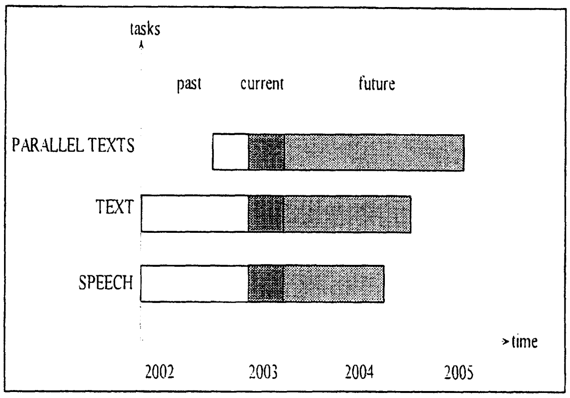

Fig. 5 Time table

a. Written Texts.

Texts in written language are collected base on time and domain criteria. Due to the wide usage of QuocNgu (national language) and political events around the 20^{th} century, we decided to collect data based on three periods: before 1945 (10%), from 1946 to1975 (40%) and from 1976 to present (50%). We classified texts into sub-domains as shown in table 5.

b. Spoken corpus

The spoken corpus is also collected according to some selective criteria such as various regions: the North (Hanoi-45%), the Center(Hue-15%), the South (Saigon-30%) and other(10%); various sources: broadcast-news(15%), telephone(10%), dialogue(40%), monologue(30%) and other (5%); various age bands: less than 10(5%), from 10 to 20(15%), from 20 to 40(65%), from 40 to 50 (10%) and older than 50(5%); sex: male (57%) and female (43%).

Table 5 Domain Criteria

Domain		Percent
Sciences		10
Historical documents		2.8
Commerce		6
World Affairs		2
Literature	Short story	12
	Novel	11
	Poems	6
	Lyrics	0.2
News	World	5
	Politics	5
	Society	8
	Culture	8
	Sports	6
	Education	5
	Health	2
	Technology & Sciences	2
	Entertainment	2
	Rural	4
Other		3

c. Parallel text corpus

A bilingual (Vietnamese-English) corpus is a part of this project, gathering several kinds of texts with various degrees of difficulty for the alignment task. The total size of the corpus is approximately 16 million words (half Vietnamese, half English). Text selection is based specific domains: novel, essay, autobiography, play, science paper, manual, dialog, etc.

d. Corpus encoding, annotating and storing

The first problem that we have to face when constructing the VnCorpus is the problem of various fonts appeared in documents. This is due to the fact that there is no standard in using fonts in documents and also due to the habit. People in the South (Saigon) usually use two byte fonts (VNI-Times, VN-Aptima, etc.) while in the North (Hanoi), they often use one byte font (.vnTimes,.vnArial, etc.). Fortunately, today we can use Unicode standard (UTF-8) to encode all Vietnamese single words. So the first step in the preprocessing stage is to convert all the documents that we have collected to Unicode fonts. Having all documents in Unicode fonts, we can exploit the available tools which originally created for English to do the preprocessing tasks such as tokenizer and sentence splitter. One of such tools that we chose is GATE [8] because it can accept documents in Unicode format. However in order to use GATE for Vietnamese, we need to define our own tokenizer rules, token types, word, number, symbol, punctuation, etc. The next stage in building this corpus is the encoding of the texts. We used SGML[7] and TEI[9] to encode texts with important information such as the boundary and part of speech of each word, sentence structure, paragraphs, sections, headings, speech turns, pausing, and para-linguistic features such as laughter in spoken texts and meta-textual information about the source or encoding of individual texts ... The last stage in creating the corpus is to add detailed descriptive information to each text, in the form of a header, and to validate the SGML structure of the whole. Header information was added to each text in the corpus from our database, giving information specific to each text, such as the author's name, or the location where a conversation was recorded. These headers are intended for use by computer programs rather than human beings, but their basic content is fairly comprehensible.

3 Some first experiences

a. Data available in the first release. In the first release of the first part of the corpus a total of some 50 million of single words taken from News [10-11] are available. Table 6 summarizes the data.

Table 6. Data available in the first release

Source		Number of texts	Num. of Sent.	Num of Single words
TEXT (News)	World	2411	289736	5208621
	Politics	2389	279814	5098172
	Society	3218	443092	8417795
	Culture	3150	438246	8287649
	Sports	2812	386482	6238346
	Education	2504	292465	5198482
	Health	927	131693	1979449
	Tech. & Sci.	1229	151284	2181848
	Entertainment	1201	167351	2079288
	Rural	2068	237892	4158897
Speech (news)	60 hours		51432	756348
Speech (Lectures)	30 hours		26891	377652
Parallel text	Dialog	6876	29812	4102180
Total		21909	2896378	49982547

b. Compound word segmentation.

Turn back to the above problem mentioned in section 1, given a sentence S with N single consecutive words $S = w_1 \ldots w_N$, we need to find a correct segmentation of compound words of the sentence. The algorithm presented here, which looks similar to the tagging problem in Chinese, includes two steps:

1. Query from the dictionary to get all possible word sequences $c_1,..,c_M$, which can be formed by the combinations of consecutive single words $w_1 \ldots w_N$ taking from the sentence S.

2. Find the word sequnces $c_1 \ldots c_L$, $1 \le L \le M$, that maximizes:

$$\left[c_1^L\right]_{opt} = \arg\max_{c_1^L, L} \left\{\Pr\left(c_1^L \middle| w_1^N\right)\right\} = \arg\max_{c_1^L, L} \left\{\Pr\left(w_1^N \middle| c_1^L\right)\Pr\left(c_1^L\right)\right\}$$

$$\approx \arg\max_{c_1^L, L} \prod_{i=1}^{L} \Pr\left(c_i \middle| c_{i-1}\right) \cdot \Pr\left(w_i \middle| c_i\right)$$

A network of HMMs was build to solve the above optimal problem, where each state is a word sequence c_i, the transition probabilities $\Pr(c_i|c_{i-1})$ are the probabilities of moving from word sequences c_{i-1} to word sequences c_i if we assume a category bigrams, and $\Pr(w_i|c_i)$ are the output probabilities.

We have tested the algorithm in a total of 1000 short sentences with 16862 isolated words. The best result is 91% of correct compound word segmentation.

c. Continuous speech recognition

In the last experience in using the VnCorpus, we have extended the syllable model [2], which has been used for recognizing isolated speech word to recognize the large vocabulary continuous speech. In this experience each isolated word was encoded by two consecutive models: consonant and syllable (for more details, see [2]). Thirty hours of broadcast news with a total of 3428 transcribed different isolated words are used to train the system. The overall of the decoding process is illustrated in Fig 5. We have tested the system on 100 sentences with a total of 1804 words. The best result is 78% of correct word recognition.

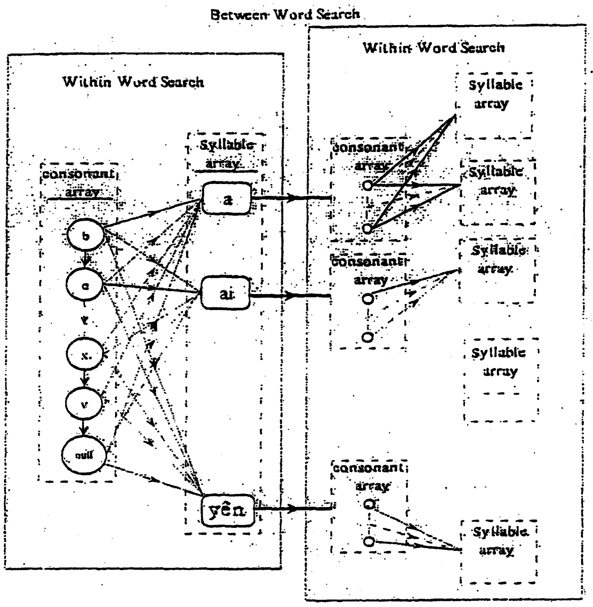

Fig. 6 Overview of the decoding process

4 Conclusion

We have proposed a framework for designing and implementing a multi-objective corpus for Vietnamese language. Some first experiences in using the corpus also mentioned. This corpus, once sufficiently extended, will be useful for training and testing NLP tools: taggers, checkers, term extractors, robust parsers, encoders, information retrieval, information extraction, machine translation, etc and spoken language processing: acoustic models, speech recognition, spoken language translation, etc.

References

[1] Hoang Phe, "Syllable Dictionary", Danang publisher, Vietnam, 1996.

[2] Vu Hai Quan, Pham Nam Trung, Nguyen Duc Hoang Ha, "A Robust Method for the recognition of Vietnamese Handwritten and Speech Recognition", ICRP2002, Quebec, Canada, 2002.

[3] Vu Hai Quan, Pham Nam Trung, Nguyen Duc Hoang Ha, "A System for Recognizing Vietnamese Document Image Based on HMM and Linguistics", ICDAR'01, pp 627-630 Seattle, USA, 2001.

[4] Vu Hai Quan, Pham Nam Trung, Nguyen Duc Hoang Ha, "Models for Vietnamese Document Image Recognition", ICISP2001, pp 484 – 491, Agadir, Morocco, 2001.

[5] Nguyen Nhu Y, "Vietnamese Dictionary", Educational Publishing House, 1997.

[6] Nguyen Tai Can, "History of Vietnamese Phonetics ", ___, ___.

[7] "The BNC Corpus", http://www.hcu.ox.ac.uk/BNC/

[8] "GATE tool", http://gate.ac.uk/

[9] "TEI Standard", http://www.tei-c.org/

[10] National Voice of Vietnam, http://www.vov.ogr.vn

[11] Laodong Newspaper, http://www.laodong.com.vn