H TÚ B O [a,b], L NG CHI MAI [a]

Tóm t t: Bài vi t này nh m gi i thi u nh ng khái ni m c b n và tình hình nghiên c u

1. M đ u

G n đây khi có d p nói chuy n v x lý ngôn ng (XLNN) và x lý ti ng Vi t (XLTV) trong công ngh t

2. Nh ng khái ni m c b n

Ti ng nói và ch vi t là hai y u t c b n nh t c a b t k ngôn ng nào. Trong s phát tri n c a công

(a) Tr c h t là các b gõ ch Vi t và thành công c a vi c đ a mã ch Vi t vào b
http://nomfoundation.org

(b) Ti p theo có th k đ n các ch ng trình nh n d ng ch Vi t in (OCR: optical character recognition), nh h Vn OCR

(c) Các ph n m m h tr vi c s d ng ngo i ng c ngoài, tiêu bi u là các t đi n song ng trên

(d) Các n l c trong vi c phát tri n các ph n m m d ch Anh-Vi t, Vi t-Anh nh các h d ch EVTRAN và VETRAN.

(e) M t lo i vi c n a là Vi t hoá các ph n m m mà g n đây tiêu bi u là k thu t Vi t hoá Windows và

Tuy liên quan đ n ti ng Vi t, không ph i t t c các vi c k trên đ u thu c v lĩnh v c x lý ngôn ng t

Để làm sáng t đi u này x *lý thông tin qua* [illegible]

Các d li u liên quan đ n ngôn ng vi t (văn b n) và nói (ti ng nói) đang d n tr nên ki u d li u chín

Gi s chúng ta có các câu sau trong các ti ng n c ngoài:

- “We meet here today to talk about Vietnamese language and speech processing.”

- “Aujourd'hui nous nous réunissons ici pour discuter le traitement de langue et de parole vietnamienne.”

- “Мы встречаемся здесь сегодня, чтобы говорить о вьетнамском языке и обработке речи.”

- “今日我々はここに集まりベトナム語処理について議論します.”

- “오늘우리는여기에모여서베트남어와발언처리에대하여의론하겠습니다 .”

Và gi s chúng ta không ai bi t c năm th ti ng trên, nh ng tò mò mu n bi t các câu đó nói gì. N u

- “Hôm nay chúng ta g p nhau ở đây để bàn v x lý ngôn ng và ti ng nói ti ng Vi t.”

N u các câu này đ c l u tr nh các t p ti ng Anh, Pháp, Nga, Nh t, Hàn và Vi t nh ta nhìn th y ở

Tuy nhiên, m t văn b n th t s (m t bài báo khoa h c ch ng h n) có th có d n hàng nghìn câu, và ta

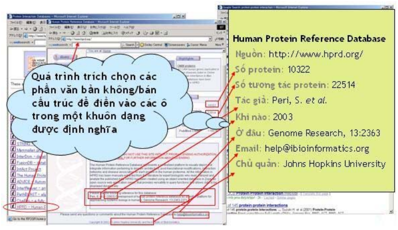
Có th hình dung ph n m m gõ ch Vi t cho phép ta tr c ti p t o ra m t t p văn b n trên máy tính (nh

1. Nh n d ng ti ng nói (speech recognition): nh n bi t và chuy n chúng thành d
2. T ng h p ti ng nói (speech synthesis): đ c m t văn b n, phân tích và chuy n thành ti ng ng
3. Nh n d ng ch vi t (optical character recognition, OCR): nh n bi t t ng ch cái và
4. D ch t đ ng (machine translation): d ch m t t p d li u văn b n trong m t ngôn ng (ti ng
5. Tóm t t văn b n (text summarization): t văn b n dài (m i trang ch ng h n) máy tóm t
6. Tìm ki m thông tin (information retrieval): t nhi u văn b n hay ti ng nói, tìm

[1]

Hình 1

7. Trích chọn thông tin (information extraction) nhiều *tập văn bản* hay tiếng nói, tìm
[2]



Hình 2

8. Phát hiện tri thức và khai phá dữ liệu văn bản (knowledge discovery and text data

Còn nhiều bài toán và công nghệ xử lý ngôn ngữ khác, như giao diện người máy bằng ngôn ngữ tự nhiên

ng dụng của công nghệ xử lý ngôn ngữ http://babelfish.altavista.com), phần mềm Nhật-Anh http://www.excite.co.jp/world/english), hay http://www.google.com/language_tools?hl=en

Có th□ phân lo□i các bài toán:

–1-3 thu□c lĩnh v□c *x□ lý ti□ng nói và x□ lý □nh* (speech and image processing),

– 4-5 thu□c lĩnh v□c *x□ lý văn b□n* (text processing),

– 6-8 thu□c lĩnh v□c *khai phá văn b□n và Web* (text and Web mining).

Phân lo□i này là t□□ng đ[illegible]

Các bài toán 1-3, 5-8 liên quan t□i *m□t ngôn ng□* “x□ lý”, trong khi (g) “x□ lý” *hai ngôn ng□ khác nhau*. Khi đ□

3. V□ s□ phát tri□n c□a x□ lý ngôn ng□ và ti□ng nói trong CNTT

Có th□ nói x□ lý ngôn ng□ t□ đ□ng trên máy tính là m□t trong nh□ng v□n đ□ khó nh□t c□a CNTT. Cái kh

M□u ch□t □ đây là b□n ch□t *ph□c t□p* c□a ngôn ng□ [illegible]

Công ngh□ ngôn ng□, nh□t là x□ lý văn b□n, v□ đ□i th□ bao g□m các b□□c (t□ng, layer) c□ b□n sau đ

1. T□ng ng□ âm (phonetic and phonological layer): Nghiên c□u v□ âm (linguistic sounds), nh□ mô hìn

2. T□ng hình thái (morphological layer): Nghiên c□u v□ các thành ph□n có nghĩa c□a t□ (word),

3. Tng ng pháp (morphological layer): Nghiên

(a) *Xác đnh t loi* (POS tagging): Xem mi t trong câu là loi gì (danh t

(b) *Xác đnh cm t* (chunking): Thí d “Ông già” là cm danh t, “đi” là c

(c) *Xác đnh quan h ng pháp* (parsing): (Ông già) (đi) (nhanh quá) là quan h ch ng

4. Tng ng nghĩa (semantic layer): Nghiên cu xác đnh nghĩa ca tng t và t hp c

5. Tng ng dng (pragmatic layer): Nghiên cu mi quan h gia ngôn ng và ng cnh s

Vic phân tích mt câu *phân tích sơ câu* trong văn (shallow parsing). Nhng *phân tích* đy đ ng pháp (full)

Nhn dng ting nói là mt quá trình nhn dng *biên vi các bin* là các đn v nhn dng, có th

– Tín hiu ting nói đưc biu din chính xác bi các giá tr ph trong mt khung thi gian ngn (shor

– Ni dung ca ting nói đưc biu din dưi dng ch vit, là mt dãy các ký hiu ng âm. Do đó ý n

– Nhn dng ting nói là mt quá trình nhn thc. Ngôn ng nói có nghĩa, do đó thông tin v ng nghĩa

Lĩnh vc nghiên cu ca nhn dng ting nói là khá rng liên quan đn nhiu ngành khác nhau, Thông

T□ng h□p ti□ng nói (text-to-speech, TTS) có [illegible]

– Ph□□ng pháp th□ nh□t đ□□c th□c hi□n d□□i d□ng các quy t□c mô t□ âm v□, □nh h□□ng l□n nhau gi□a

– Ph□□ng pháp th□ hai l□u gi□ nh□ng đ□n v□ âm c□ b□n, bi□n đ□i đ□n v□ âm c□ b□n và đ□ng th□i t□o r

Trong khi x□ lý ti□ng Vi□t còn đang ch□p ch□ng nh□ng b□□c đi đ□u, các nghiên c□u và □ng d□ng v□ x□

Các lý thuy□t n□n t□ng đ□□c xây d□ng trong các năm 1940 và 1950: [illegible] có □nh h□□ng sâu s□c đ□n x□ l

Hai nhánh tách r□i nhau trong các năm 1950 đ□n năm 1970: [illegible] (a) ki□u hình th□c (symbolic paradigm) ch

B□n ki□u x□ lý ngôn ng□ ph□ bi□n trong th□p k□ 70 đ□n gi□a th□p k□ 80

(i) *các mô hình ng□u nhiên* đóng m□t vai trò l□n, tiêu bi□u là các mô hình Markov □

(ii) *ki□u d□a trên logic* (logic-based paradigm);

(iii) *hi□u ngôn ng□ t□ nhiên* (natural language understanding);

(iv) *mô hình hoá các cu□c đ□i tho□i* (discourse modeling).

Cũng trong th□i gian này xu□t hi□n các bài toán và tài nguyên chu□n (Penn Treebank, WordNet, MUC, e

Ch□ nghĩa kinh nghi□m và các mô hình có th□ m□nh thái c□a th□p k□ 80 đ□n gi□a th□p k□ 90

X□ lý văn b□n và ti□ng nói g□p nhau trong m□□i năm qua: Đây là giai đo□n ti□n b□ v□i mô hình th□ng kê và ti□p c□n d□a vào c

Nh□ đã trình bày s□ b□□ trên, x□ lí ngôn ng□ là m□t vi□c khó, ph□c t□p, ch□ có th□ làm lâu dài theo nhi□

Trên th□ gi□i, nhi□u t□ ch□c qu□c t□, nhi□u hi□p h□i x□ lí ngôn ng□ t□ nhiên đã đ□□c thành l□p v□i các h http://www.ldc.upenn.edu).

Nhi□u chính ph□ đã đ□u t□ l□n cho x□ lý ngôn ng□ trong CNTT (M□, Nh□t b□n, Trung qu□c, Singapore, http://www.tcllab.org).

Là ng□□i đi sau trong lĩnh v□c x□ lí ngôn ng□ … *Biết, hiểu và khai thác, vận dụng các công nghệ, các công cụ …*

4. Tình hình và nh□ng v□n đ□ chính trong x□ lý ti□ng Vi□t

Hãy th□ nhìn l□i tình hình c□a chúng ta. Ngoài nh□ng vi□c đã làm và b□□c đ□u làm đ□□c k□ □ ph□n đ□u

Bên ngoài Vi□t Nam, cũng có nh□ng n□ l□c v□ x□ lý ti□ng Vi□t, nh□ nhóm d□ch Anh-Vi□t c□a ti□n s□ Ph

Ngoài nh□ng k□t qu□ ban đ□u, sau đây có th□ là m□t vài đ□c đi□m chính v□ ho□t đ□ng x□ lý ngôn ng□ c

- Th□□ng t□p trung vào làm các s□n ph□m cho ng□□i dùng cu□i … s□n ph□m d□ch máy, m□t lo

- Ít các nghiên c□u n□n t□ng, x□ lý ng□ … phát tri□n “h□ t□ng c□ s□” và tài nguyên: t□ đi□n (

- Ph□n đông là các nghiên c□u ng□n h□n và đ□im hi□u m□c tài thc s□, và đi□u ki□n. Đa s□ m□i

- R□t có th□ nhi□u nhóm đã b□t đ□u ch□a đ□y đ□, ho□c ti□n hành công vi□c khi còn thi□u ki□n th□c. Đo v□

- Đáng băn khoăn h□n c□ là các c□ liên k□t, chúng ta s□ ti□ng Vi□t, và □ phát tri□n t□ th□ có k□ ho

5. K□t lu□n

Chúng ta ai cũng bi□t x□ lý ti□ng Vi□t là m□t vi□c khó và lâu dài, đòi h□i nhi□u công s□c c□a nhi□u ng□ i. Ngoài

1. Xây d□ng và phát tri□n m□t x□ lý ti□ng Vi□t và tìm ki□m thông tin trên Internet b□

2. Xây d□ng các công c□ và ngu□n tài nguyên thi□t y□u, th□c hi□n m□c tiêu 1 c□a c

M□c dù vi□c ph□i t□o các s□n ph□m cho ng□□i dùng cu□i là c□p bách và đích cu□i cùng ta c□n đ□n, ch

L□i cám □n

Xin chân thành cám □n các ý ki□n th□o lu□n, góp ý cho b□n th□o tài li□u này c□a các đ□ng nghi□p: Ngô

Tài li□u tham kh□o chính

Allen, J. (1994). Natural Language Understanding. The Benjamin/Cummings Publishing Co.

Baeza-Yates, R., Ribeiro-Neto, B. (1999). Modern Information Retrieval, Addison Wesley.

Bao, H.T., Thang, N.T., Chien, N.P., Mai, L.C. (2001). Towards a Practical Framework for Vietnamese N

Bao, H.T., Funakoshi, K. (1998). Information Retrieval Using Rough Sets, Journal of Japanese Society

Bao, H.T., Binh, N.N. (2002). Nonhierarchical Document Clustering by a Tolerance Rough Set Model, In

Bao, H.T., Tuan, N.A., Son, N.C. (2003). Issues in Construction of a Vietnamese Machine Tractable Dic

Berry, M.W. (2004). Survey of Text Mining: Clustering, Classification, and Retrieval, Springer.

Chakrabarti, S. (2003). Mining the Web, Morgan Kaufmann Publishers.

Cohen, W., McCallum, A. (2003). Information Extraction from the World Wide Web, Tutorial in ACM Co

Cole, R., Mariani, J., Uszkoreit, H., Varile, G., Zaenen, A., Zampolli, A., Zue, V. (1997). Language Tech

Dale, R., Moisl, H., Somers, H. (2000). Handbook of Natural Language Processing, Marcel Dekker.

Dien, D., Kiem, H., Toan, N.V. (2001). Vietnamese Word Segmentation, Proceedings of the Sixth Natur

Dien, D. (2002). Building a Training Corpus for Word Sense Disambiguation in English-to-Vietnamese M

Dien, D., Kiem, H. (2003). POS-Tagger for English-Vietnamese Bilingual Corpus, Proceedings of HLT-N

Door, B.J., Jordan, P.W., Benoit, J.W. (2000). A Survey of Current Paradigms in Machine Translation.

EDR Electronic Dictionary Technical Guide (1993). Japan Electronic Dictionary Research Institute.

Jelinek, F. (1998). Statistical Methods for Speech Recognition. The MIT Press.

Jurafsky, D., Martin, J. H. (2000). Speech and Language Processing. An Introduction to Natural Langua

Hieu, P.X., Horiguchi, S., Bao, H.T. (2005). Conditional Models for Automatic Data Integration from the

Huong, L.T. (2004). Investigation into an Approach to Automatic Text Summarisation, Ph.D. dissertation

Huyen, N.T.M., Laurent Romary, Luong, V.X. (2003). A Case Study in POS Tagging of Vietnamese Tex

Khang, B.H. et al. (2004). Báo cáo T ng k t Khoa h c và K thu t Đ tài Nghiên c u Phát tri n Công n

Mani, I., Maybury, M.T. (1999). Advanced in Automatic Text Summarization, The MIT Press.

Manning C. D., Schutze H. (1999). Foundations of Statistical Natural Language Processing. MIT Press.

Minh, N.L., Shimazu, A., Horiguchi, S., Bao, H.T. (2004). Example-Based Sentence Reduction Using Hi

Minh, N.L., Shimazu, A., Horiguchi, S., Bao, H.T., Fukushi, M. (2004). Probabilistic Sentence Reduction

Minh, N.L. (2004). Statistical Machine Learning Approaches to Cross-language Text Summarization, Ph

Nagao, M. (1989). Machine translation: how far can it go? Oxford University.

Oracle Text – An Oracle White Paper (2001) http://www.oracle.com/technology/products/text/pdf/text_bwp.pdf

Sirmakessis S. (2004). Text Mining and Its Applications, Springer.

[a,b] Vi n Công ngh Thông tin, [b] Vi n Khoa h c và Công ngh T

[1] Chú thích c a Vietlex.
[2] Chú thích c a Vietlex.