Về xử lý tiếng Việt trong công nghệ thông tin

Hồ Tú Bảo [a,b], Lương Chi Mai [a]
[a]Viện Công nghệ Thông tin, [b]Viện Khoa học và Công nghệ Tiên tiến Nhật bản

Tóm tắt: Bài viết này nhằm giới thiệu những khái niệm cơ bản và tình hình nghiên cứu về xử lý ngôn ngữ nói chung, cũng như những nội dung và khó khăn trong xử lý tiếng Việt (văn bản và tiếng nói). Bài viết này có thể được dùng như một tài liệu tham khảo cho các nhà quản lý khoa học và những người làm nghiên cứu khoa học – công nghệ không chuyên về lĩnh vực xử lý ngôn ngữ.

1. Mở đầu

Gần đây khi có dịp nói chuyện về xử lý ngôn ngữ (XLNN) và xử lý tiếng Việt (XLTV) trong công nghệ thông tin (CNTT) với một số nhà quản lý khoa học và công nghệ hoặc một số đồng nghiệp, chúng tôi thấy có sự khác nhau giữa nhiều người về cách hiểu một số khái niệm cũng như giữa những nhìn nhận về tình hình nghiên cứu-phát triển trong lĩnh vực này. Điều này cũng tự nhiên, tự nhiên như hầu hết chúng ta không thật rõ về bệnh tim, hay không rõ protein được tổng hợp ra như thế nào. Khi chuẩn bị dự án về xử lý tiếng Việt, chúng tôi bỗng thấy cần giải thích cho nhiều người không làm chuyên môn về xử lý ngôn ngữ rõ hơn về các câu chuyện của lĩnh vực này. Và thay vì viết ngay đề cương, chúng tôi bắt đầu các việc của dự án bằng bài viết này.

2. Những khái niệm cơ bản

Tiếng nói và chữ viết là hai yếu tố cơ bản nhất của bất kỳ ngôn ngữ nào. Trong sự phát triển của công nghệ thông tin (CNTT) ở Việt Nam, một số việc liên quan đến "tiếng Việt" đã được làm và ít nhiều có kết quả ban đầu:

(a) Trước hết là các bộ gõ chữ Việt và thành công của việc đưa được bộ mã chữ Việt vào bảng mã Unicode, cũng như việc chọn Unicode cho bộ mã chuẩn tiếng Việt (nhân đây cũng xin nói thêm, do chưa ý thức về chuẩn, rất nhiều cán bộ CNTT, nhiều cơ quan nhà nước vẫn chưa chịu đổi thói quen cũ để dùng bộ mã chuẩn Unicode, một việc rất quan trọng của xử lý tiếng Việt). Bảo tồn chữ Nôm trên máy tính cũng là một việc đầy nỗ lực và nhiều ý nghĩa được nhiều người theo đuổi lâu nay, cần được nhà nước tiếp tục ủng hộ lâu dài (http://nomfoundation.org).

(b) Tiếp theo có thể kể đến các chương trình nhận dạng chữ Việt in (OCR: optical character recognition), như hệ VnDOCR của Viện Công nghệ Thông tin, Viện Khoa học và Công

nghệ Việt Nam. Các chương trình nhận dạng chữ in nhằm *chuyển* các tài liệu in trên giấy thành các tài liệu điện tử (dưới dạng các tệp văn bản trên máy tính).

(c) Các phần mềm hỗ trợ việc sử dụng tiếng nước ngoài, tiêu biểu là các từ điển song ngữ trên máy tính, thí dụ như các từ điển điện tử của Lạc Việt đã được dùng rộng rãi trên máy tính để tra cứu từ Anh-Việt, Việt-Anh. Điều ta cần phân biệt là các từ điển điện tử này dành cho con người sử dụng, khác với từ điển điện tử dành cho máy tính sử dụng trong xử lý ngôn ngữ tự nhiên (sẽ được đề cập ở phần sau).

(d) Các nỗ lực trong việc làm các phần mềm dịch Anh-Việt,Việt-Anh, chẳng hạn như các hệ dịch EVTRAN và VETRAN.

(e) Một loại việc nữa là Việt hóa các phần mềm mà gần đây tiêu biểu là kết quả Việt hóa Windows và Microsoft Office của Microsoft. Việc này có thể xem như việc "dịch" các thông báo tiếng Anh *cố định* trong các phần mềm thành các thông báo tiếng Việt.

Tuy liên quan đến tiếng Việt, không phải tất cả các việc kể trên đều thuộc về lĩnh vực xử lý ngôn ngữ tự nhiên nói chung và xử lý tiếng Việt nói riêng theo nghĩa thông thường trong CNTT, vốn chủ yếu nhằm vào những vấn đề liên quan đến xử lý văn bản (text) và tiếng nói (speech) [Jurafsky and Martin, 2000].

Để làm sáng tỏ điều này ta thử xem xét lại khái niệm "xử lý thông tin", một khái niệm cốt lõi của công nghệ thông tin và là khái niệm rộng hơn "xử lý ngôn ngữ". Về bản chất, *xử lý thông tin là quá trình biến đổi dữ liệu từ dạng này thành dạng khác để có thể thu được thông tin và tri thức*. Trong giai đoạn đầu, CNTT tập trung vào các dữ liệu dạng số, biểu diễn bởi các dạng được cấu trúc (structured) như các vectơ (vector) hay bảng biểu (tables). Trong hơn nửa thế kỷ phát triển, CNTT dần dần "xử lý" nhiều kiểu dữ liệu khác, như hình ảnh (image), âm thanh (voice, speech), văn bản (text), kí hiệu hình thức (symbols), đồ thị (graph),… và gần đây là nhiều kiểu dữ liệu phức tạp như dữ liệu sinh học (genomic data). Phương pháp xử lý cũng ngày càng phong phú, từ tính toán (computing) đến suy luận (reasoning), và nhiều kiểu khác nữa. Xử lý ngôn ngữ chính là xử lý thông tin khi đầu vào là "dữ liệu ngôn ngữ" (dữ liệu cần biến đổi), tức dữ liệu "văn bản" hay "tiếng nói".

Các dữ liệu liên quan đến ngôn ngữ viết (văn bản) và nói (tiếng nói) đang dần trở nên kiểu dữ liệu chính con người có và lưu trữ dưới dạng điện tử. Đặc điểm chính của các kiểu dữ liệu này là không có cấu trúc hoặc nửa cấu trúc (non-structured hoặc semi-structured) và chúng không thể lưu trữ trong các khuôn dạng cố định như các bảng biểu. Theo đánh giá của công ty Oracle, hiện có đến 80% dữ liệu không cấu trúc trong lượng dữ liệu của loài người đang có [Oracle Text]. Với sự ra đời và phổ biến của Internet, của sách báo điện tử, của máy tính cá nhân, của viễn thông, của thiết bị âm thanh, … người người ai cũng có thể tạo ra dữ liệu văn bản hay tiếng nói. Vấn đề là làm sao ta có thể xử lý chúng, tức chuyển chúng từ các dạng ta chưa hiểu được thành

các dạng ta có thể hiểu và giải thích được, tức là ta có thể tìm ra thông tin, tri thức hữu ích cho mình.

Giả sử chúng ta có các câu sau trong các tiếng nước 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 встрачаемся 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."
- "오늘 우리는 여기에 모여서 베트남어와 발언처리에 대하여 의론하겠습니다."

Và giả sử chúng ta không ai biết cả năm thứ tiếng trên, nhưng tò mò muốn biết các câu đó nói gì. Nếu có ai đó dịch, hoặc có một chương trình máy tính dịch (biến đổi) chúng ra tiếng Việt, ta sẽ hiểu nghĩa các câu trên đều là:

- "Hôm nay chúng ta gặp nhau ở đây để bàn về xử lý ngôn ngữ và tiếng nói tiếng Việt."

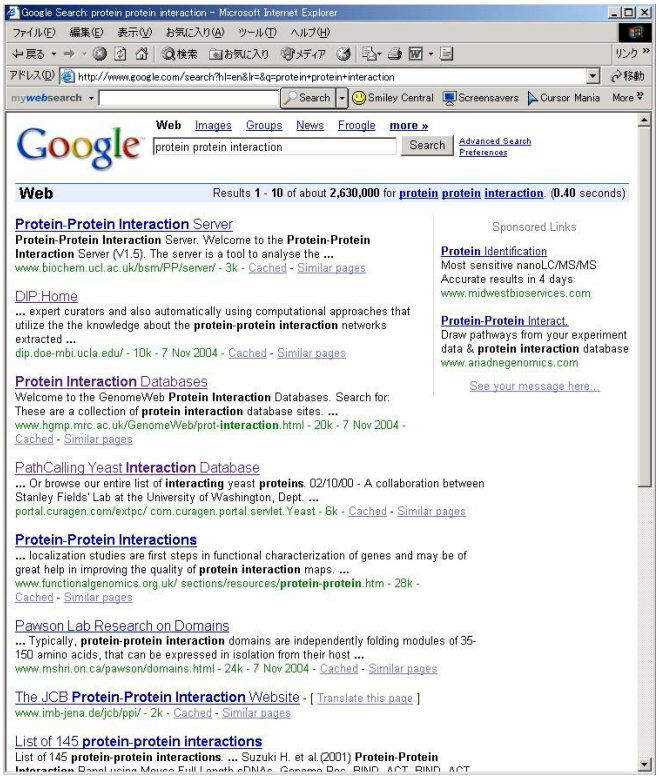
Nếu các câu này được lưu trữ như các tệp tiếng Anh, Pháp, Nga, Nhật, Hàn và Việt như ta nhìn thấy ở trên, ta có các dữ liệu "văn bản". Nếu ai đó đọc các câu này, ghi âm lại, ta có thể chuyển chúng vào máy tính dưới dạng các tệp các tín hiệu (signal) "tiếng nói". Tín hiệu sóng âm của hai âm tiết tiếng Việt có thể nhìn thấy như sau

Tuy nhiên, một văn bản thật sự (một bài báo khoa học chẳng hạn) có thể có đến hàng nghìn câu, và ta không phải có một mà hàng triệu văn bản. Web là một nguồn dữ liệu văn bản khổng lồ, và cùng với các thư viện điện tử − khi trong một tương gần các sách báo xưa nay và các nguồn âm thanh được chuyển hết vào máy tính (chẳng hạn bằng các chương trình nhận dạng chữ, thu nhập âm thanh, hoặc gõ thẳng vào máy) − sẽ sớm chứa hầu như toàn bộ kiến thức của nhân loại. Vấn đề là làm sao "xử lý" (chuyển đổi) được khối dữ liệu văn bản và tiếng nói khổng lồ này qua dạng khác để mỗi người có được thông tin và tri thức cần thiết từ chúng? Điều này càng quan trọng khi đa số nguồn tri thức quý giá này lại bằng tiếng nước ngoài và đa số người Việt chúng ta còn ít nghe hay đọc được chúng.

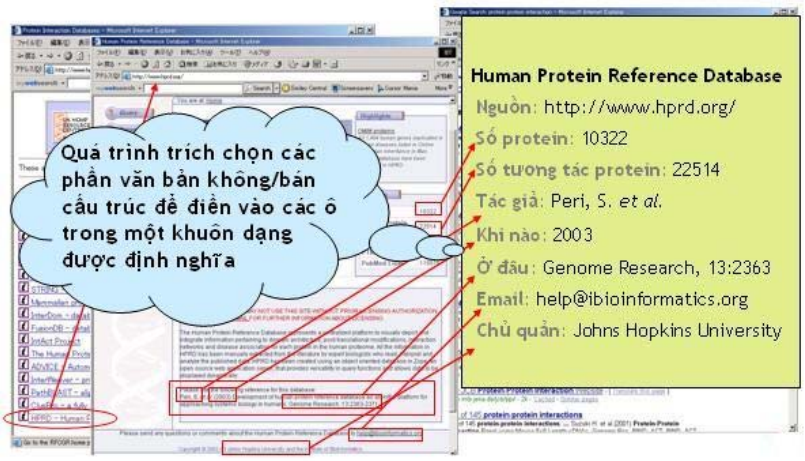
Có thể hình dung phần mềm gõ chữ Việt cho phép ta trực tiếp tạo ra một tệp văn bản trên máy tính (như chúng tôi đang gõ máy tính để viết bài này), còn chương trình nhận dạng chữ

VnDOCR cho phép ta biến đổi một văn bản in trên giấy thành một tệp văn bản trên máy tính. Tuy nhiên, các sản phẩm trên vẫn chưa động chạm đến cốt lõi của xử lý ngôn ngữ. Theo nghĩa "xử lý ngôn ngữ" nêu ở trên − tức biến đổi dữ liệu ngôn ngữ − có thể nêu ra một số bài toán tiêu biểu của xử lý ngôn ngữ với các mức độ khác nhau về xử lý và sử dụng ngôn ngữ tự nhiên của con người:

1. Nhận dạng tiếng nói (speech recognition): từ sóng tiếng nói, nhận biết và chuyển chúng thành dữ liệu văn bản tương ứng [Jelinek, 1998], [Jurafsky and Martin, 2000].
2. Tổng hợp tiếng nói (speech synthesis): từ dữ liệu văn bản, phân tích và chuyển thành tiếng người nói [Jelinek, 1998], [Jurafsky and Martin, 2000].
3. Nhận dạng chữ viết (optical character recognition, OCR): từ một văn bản in trên giấy, nhận biết từng chữ cái và chuyển chúng thành một tệp văn bản trên máy tính.
4. Dịch tự động (machine translation): từ một tệp dữ liệu văn bản trong một ngôn ngữ (tiếng Anh chẳng hạn), máy tính dịch và chuyển thành một tệp văn bản trong một ngôn ngữ khác (tiếng Việt chẳng hạn) [Dorr et al., 2000], [Nagao, 1989].
5. Tóm tắt văn bản (text summarization): từ một văn bản dài (mười trang chẳng hạn) máy tóm tắt thành một văn bản ngắn hơn (một trang) với những nội dung cơ bản [Mani and Maybury, 1999]
6. Tìm kiếm thông tin (information retrieval): từ một nguồn rất nhiều tệp văn bản hay tiếng nói, tìm ra những tệp có nội dung liên quan đến một vấn đề (câu hỏi) ta cần biết (hay trả lời) [Baeza-Yates and Ribeiro-Neto, 1999].. Điển hình của công nghệ này là *Google*, một hệ tìm kiếm thông tin trên Web, mà hầu như chúng ta đều dùng thường xuyên. Cần nói thêm rằng mặc dù hữu hiệu hàng đầu như vậy, Google mới có khả năng cho chúng ta tìm kiếm câu hỏi dưới dạng các từ khóa (keywords) và luôn "tìm" cho chúng ta rất nhiều tài liệu không liên quan, cũng như rất nhiều tài liệu liên quan đã tồn tại thì Google lại tìm không ra. Hình bên chỉ ra một màn hình của Google đưa ra các trang Web liên quan đến câu hỏi "protein-protein interaction"



7. Trích chọn thông tin (information extraction): từ một nguồn rất nhiều tệp văn bản hay tiếng nói, tìm ra những *đoạn bên trong* một số tệp liên quan đến một vấn đề (câu hỏi) ta cần biết hay trả lời. Hình bên phải minh họa một kết quả trích chọn thông tin với cùng câu hỏi "protein-protein interaction". Một hệ trích chọn thông tin có thể "lần" vào từng trang Web liên quan, phân tích bên trong và trích ra các thông tin cần thiết, nói gọn trong tiếng Anh để phân biệt với tìm kiếm thông tin là "find things but not pages" [Cohen and McCallum, 2003].



8. Phát hiện tri thức và khai phá dữ liệu văn bảm (knowledge discovery and text data mining): Từ những nguồn rất nhiều văn bản thậm chí hầu như không có quan hệ với nhau, tìm ra được những tri thức trước đấy chưa ai biết. Đây là một vấn đề rất phức tạp và đang ở giai đoạn đầu của các nghiên cứu trên thế giới [Berry, 2004], [Sirmakessis, 2004].

Còn nhiều bài toán và công nghệ xử lý ngôn ngữ khác, như giao diện người máy bằng ngôn ngữ tự nhiên, các hệ hỏi đáp, các hệ sinh ra ngôn ngữ, …

Ứng dụng của công nghệ xử lý ngôn ngữ hết sức phong phú. Có thể lấy vài thí dụ gần đây. Tin trên Internet ngày 21/4/2005, hãng Samsung đưa ra thị trường điện thoại di động P207 có thể nhận biết được các câu nói đơn giản của người sử dụng điện thoại di động như "Hãy gọi cho tôi" hay "Tôi sẽ gọi lại", rồi chuyển chúng thành dạng văn bản (tin nhắn) cho người dùng điện thoại di động gửi nhắn tin. Đây là một ứng dụng của QuickPhrase trong VoiceSignal Technologies, tất nhiên là cho tiếng Hàn. Ta có thể hình dung đây là việc ghi lại tín hiệu một câu nói đơn giản, so sánh tín hiệu này với tín hiệu của một rất nhiều câu nói đã được ghi trước rồi chọn câu gần nhất (http://www.vnexpress.net/Vietnam/Vi-tinh/San-pham-moi/2005/04/ 3B9D D713/). Tin ngày 22/4/2005 (http://www2.tuoitre.com.vn/Tianyon/Index.aspx?ArticleID=75496&ChannelID=17) cho biết công ty CombiWith (Nhật) công bố sắp tung ra các bức ảnh biết nói để cho những người cô đơn ở Nhật có thể có ảnh của người thân biết nói với chính giọng của họ. Có thể hình dung đây chính là việc tổng hợp tiếng nói của người trong ảnh dựa trên việc học giọng nói từ rất nhiều mẫu câu nói của người đó. Tất nhiên, đây lại là tiếng Nhật vì công nghệ cho tiếng Việt sẽ rất khác. Cũng vậy, có rất nhiều phần mềm dịch tự động trên Web, như Babel Fish Translation của AltaVista (http://babelfish.altavista.com/), dịch Anh-Nhật, Nhật-Anh bởi Excite bản tiếng Nhật

(http://www.excite.co.jp/world/english/), hay Language Tools dịch nhiều thứ tiếng của Google (http://www.google.com/language_tools?hl=en).

Có thể phân loại các bài toán:

- 1-3 thuộc lĩnh vực *xử lý tiếng nói và xử lý ảnh* (speech and image processing),
- 4-5 thuộc lĩnh vực *xử lý văn bản* (text processing),
- 6-8 thuộc lĩnh vực *khai phá văn bản và Web* (text and Web mining).

Phân loại này là tương đối, vì các lĩnh vực trên có rất nhiều phần chung nhau. Về bản chất, xử lý tiếng nói dựa hay ảnh trên các kỹ thuật *phân tích và nhận dạng tín hiệu* (signal processing and recognition). Xử lý văn bản cũng như khai phá văn bản và Web lại dựa trên các kỹ thuật của *xử lý ngôn ngữ tự nhiên* (natural language processing hay computational linguistics) mà quan trọng là việc *hiểu* (understanding) và *dùng tri thức về ngôn ngữ* ở các mức độ khác nhau [Jurafsky and Martin, 2000]. Nếu các bài toán 1-5 có đối tượng xử lý là *một tệp* văn bản hay tiếng nói, thì các bài toán 6-8 có đối tượng xử lý là một tập hợp *rất nhiều tệp* văn bản hay tiếng nói. Cần nhấn mạnh thêm là do sự phát triển của Internet, việc tìm kiếm và trích chọn thông tin, phát hiện tri thức từ các cơ sở dữ liệu lớn là các nội dung thời sự và xu thế của phát triển CNTT trên thế giới [Berry, 2004], [Chakrabarti, 2003], [Cole et al., 1997], [Sirmakessis, 2004].

Các bài toán 1-3, 5-8 liên quan việc "xử lý" *một ngôn ngữ*, trong khi (g) "xử lý" *hai ngôn ngữ khác nhau*. Khi đầu vào hay đầu ra của các bài toán 1-8 là tiếng Việt, ta có các vấn đề của *xử lý tiếng Việt*.

3. Về sự phát triển của xử lý ngôn ngữ và tiếng nói trong CNTT

Có thể nói xử lý ngôn ngữ tự động trên máy tính là một trong những vấn đề khó nhất của CNTT. Cái khó nằm ở chỗ làm sao cho máy được hiểu ngôn ngữ con người, từ việc hiểu nghĩa từng từ trong mỗi hoàn cảnh cụ thể, đến việc hiểu nghĩa một câu, rồi hiểu cả văn bản. Ta lấy lại thí dụ của Marvin Minsky (1992), một cây đa cây đề của ngành trí tuệ nhân tạo (artificial intelligence): "Xét từ "sợi dây" chẳng hạn. Ngày nay không một máy tính nào có thể hiểu nghĩa từ này như con người. Còn chúng ta ai cũng biết có thể dùng sợi dây để kéo một vật, nhưng không thể đẩy một vật bằng sợi dây này. Ta có thể gói một gói hàng hoặc thả diều bằng một sợi dây, nhưng không thể ăn sợi dây. Trong vài phút, một đứa trẻ nhỏ có thể chỉ ra hàng trăm cách dùng hoặc không dùng một sợi dây, nhưng không máy tính nào có thể làm việc này."

Mấu chốt ở đây là bản chất phức tạp của ngôn ngữ của con người, đặc biệt là sự đa nghĩa và nhập nhằng nghĩa của ngôn ngữ. Thêm nữa, có một khác biệt sâu sắc nữa là con người ngầm

hiểu và dùng quá nhiều *lẽ thường* (common sense) trong ngôn ngữ, như biết "lửa" thì nóng còn "chim" thì biết bay, hay sợi dây thì không dùng để đẩy hay khều các vật, trong khi rất khó làm cho máy hiểu các lẽ thường này.

Công nghệ ngôn ngữ, nhất là xử lý văn bản, về đại thể bao gồm các bước (tầng, layer) cơ bản sau đây [Allen, 1994], [Jurafsky and Martin, 2000]:

1. Tầng ngữ âm (phonetic and phonological layer): Nghiên cứu về ngữ âm (linguistic sounds), như mô hình hóa việc các từ trong cách nói thông thường được phát âm thế nào, về bản chất thanh điệu, ngôn điệu, ngữ điệu (prosody, intonation), trường độ âm tiết, độ nhấn, biến thanh, …
2. Tầng hình thái (morphological layer): Nghiên cứu về các thành phần có nghĩa của từ (word), như từ được tạo ra bởi các hình vị (morphemes) và từ được tách ra trong một câu thế nào. Thí dụ, từ "tiếng Việt" tạo thành từ "t-iế-ng V-iệ-t", còn ngữ (phrase) "xử lý tiếng Việt" tạo thành gồm hai từ "xử lý" và "tiếng Việt". Trong tiếng Việt, một bài toán quan trọng là phân tách từ (word segmentation). Một thí dụ quen thuộc là câu "Ông già đi nhanh quá" có thể phân tách thành (Ông già) (đi) (nhanh quá) hoặc (Ông) (già đi) (nhanh quá) hoặc những cách khác nữa.
3. Tầng ngữ pháp (morphological layer): Nghiên cứu các quan hệ cấu trúc giữa các từ, xem các từ đi với nhau thế nào để tạo ra câu đúng. Quá trình này thường được cụ thể trong các bước cơ bản sau:
   (a) *Xác định từ loại* (POS tagging): Xem mỗi từ trong câu là loại gì (danh từ, động từ, giới từ, …). Trong thí dụ trên, có thể "Ông già" là danh từ, "đi" là động từ, "nhanh" là trạng từ, và "quá" là thán từ.
   (b) *Xác định cụm từ* (chunking): Thí dụ "Ông già" là cụm danh từ, "đi" là cụm động từ, "nhanh quá" là cụm trạng từ. Như vậy câu trên có hai phân tích (Ông già) (đi) (nhanh quá) hoặc (Ông) (già đi) (nhanh quá).
   (c) *Xác định quan hệ ngữ pháp*(parsing): (Ông già) (đi) (nhanh quá) là quan hệ chủ ngữ-vị ngữ-trạng ngữ.
4. Tầng ngữ nghĩa (semantic layer): Nghiên cứu xác định nghĩa của từng từ và tổ hợp của chúng để tạo nghĩa của câu. Thí dụ trong phân tích (Ông già) (đi) (nhanh quá), động từ "đi" có thể có nghĩa "bước đi", hay "chết" hay "điều khiển" (khi đánh cờ), … và tương ứng ta có các nghĩa khác nhau của câu.
5. Tầng ngữ dụng (pragmatic layer): Nghiên cứu mối quan hệ giữa ngôn ngữ và ngữ cảnh xử dụng ngôn ngữ (context-of-use). Ngữ dụng như vậy nghiên cứu việc ngôn ngữ được dùng để nói về người và vật như thế nào.

Việc phân tích một câu nói hay một câu trong văn bản ở các tầng từ ngữ âm (1) đến tầng ngữ pháp (3) gọi là *phân tích sơ bộ* (shallow parsing). Nếu phân tích thêm cả tầng ngữ nghĩa (từ (1) đến (4)) ta sẽ có *phân tích đầy đủ* (fully parsing). Trong các vấn đề của xử lí ngôn ngữ, có vấn đề cần đến phân tích đầy đủ (như dịch tự động), nhưng cũng có những vấn đề chỉ với phân tích sơ bộ cũng có thể đã xử lí được (như tìm kiếm thông tin, phân tích văn bản cho tổng hợp tiếng nói, mô hình ngôn ngữ trong nhận dạng tiếng nói…).

Nhận dạng tiếng nói là một quá trình nhận dạng mẫu, với các mẫu là các đơn vị nhận dạng, có thể là các từ hoặc các âm vị. Khó khăn cơ bản của bài toán này là tiếng nói luôn biến thiên theo thời gian và có sự khác biệt lớn giữa tiếng nói của những người nói khác nhau, tốc độ nói, ngữ cảnh và môi trường âm học khác nhau. Các nghiên cứu về nhận dạng tiếng nói dựa trên *ba nguyên tắc cơ bản*:

- Tín hiệu tiếng nói được biểu diễn chính xác bởi các giá trị phổ trong một khung thời gian ngắn (short-term amplitude spectrum). Nhờ vậy ta có thể trích ra các đặc điểm tiếng nói từ những khoảng thời gian ngắn và dùng các đặc điểm này làm dữ liệu để nhận dạng tiếng nói.
- Nội dung của tiếng nói được biểu diễn dưới dạng chữ viết, là một dãy các ký hiệu ngữ âm. Do đó ý nghĩa của một phát âm được bảo toàn khi chúng ta phiên âm phát âm thành dãy các ký hiệu ngữ âm.
- Nhận dạng tiếng nói là một quá trình nhận thức. Ngôn ngữ nói có nghĩa, do đó thông tin về ngữ nghĩa (semantics) và ngữ dụng (pragmatics) có giá trị trong quá trình nhận dạng tiếng nói, đặc biệt khi thông tin về âm học là không rõ ràng.

Lĩnh vực nghiên cứu của nhận dạng tiếng nói là khá rộng liên quan đến nhiều ngành khác nhau, như xử lý tín hiệu số (digital signal proccessing), vật lý hay âm học (acoustic), nhận dạng mẫu, lý thuyết thông tin và khoa học máy tính (information and computer science theory), ngôn ngữ học (linguistics), sinh lý học (physiology), tâm lý học ứng dụng (applied psychology). Các hệ thống nhận dạng tiếng nói có thể được phân chia thành hai loại khác nhau: hệ thống nhận dạng từ rời rạc và hệ thống nhận dạng từ liên tục. Trong hệ thống nhận dạng tiếng nói liên tục, người ta lại phân biệt hệ thống nhận dạng có kích thước từ điển nhỏ và hệ thống nhận dạng với kích thước từ điển trung bình hoặc lớn.

Tổng hợp tiếng nói (text-to-speech, TTS) có mục tiêu ngược với mục tiêu của nhận dạng tiếng nói. Kiến trúc của một hệ thống TTS giống như kiến trúc đọc chữ của con người, bao gồm một môđun xử lý ngôn ngữ tự nhiên (bộ tiền xử lý nhằm tổ chức các câu thành danh sách, bộ phân tích hình thái, bộ phân tích ngữ cảnh, bộ phân tích câu cú pháp, ngôn điệu, …), có khả năng sinh ra phiên âm phù hợp với cách phát âm của quá trình đọc văn bản cùng với ngữ điệu, ngôn điệu;

và một môđun xử lý tín hiệu số, môđun này chuyển thông tin tượng trưng nhận được thành tiếng nói (môđun letter-to-sound và môđun sinh ra ngôn điệu). Khi hai khối xử lý ngôn ngữ tự nhiên và xử lý tín hiệu số được định nghĩa rõ ràng, việc nghiên cứu về hai quá trình có thể được thực hiện riêng rẽ, độc lập với nhau. Khối xử lý tín hiệu số phải xét đến các hạn chế phát âm, vì sự biến đổi ngữ âm (phần động, chuyển tiếp giữa các âm) là quan trọng đối với việc hiểu lời nói hơn là các phần tĩnh của lời nói. Tổng hợp tiếng nói có thể đạt được cơ bản theo hai phương pháp thuộc về hai trường phái tổng hợp tiếng nói có nội dung và mục tiêu khác nhau:

- Phương pháp thứ nhất được thực hiện dưới dạng các quy tắc mô tả âm vị, ảnh hưởng lẫn nhau giữa các âm vị khi phát ra một âm (tổng hợp bằng qui luật).
- Phương pháp thứ hai lưu giữ những đơn vị âm cơ bản, biến đổi đơn vị âm cơ bản và đồng thời tạo ra cơ sở dữ liệu tiếng nói, sử dụng chúng như là các đơn vị âm học cơ bản để tạo thành lời nói (phương pháp tổng hợp theo xích chuỗi).

Trong khi xử lý tiếng Việt còn đang chập chững những bước đi đầu, các nghiên cứu và ứng dụng về xử lý ngôn ngữ nói chung trên thế giới và nhiều nước khác đã có một lịch sử hơn nửa thế kỷ, đã trải qua nhiều giai đoạn, và điều quan trọng hơn cả là nhiều con đường và cách thức xử lí ngôn ngữ đã được trải nghiệm và thừa nhận. Lịch sử xử lý ngôn ngữ có thể được chia ra các giai đoạn như sau [Jurafsky and Martin, 2000]:

Các lý thuyết nền tảng được xây dựng trong các năm 1940 và 1950: Hai kiểu mô hình nền tảng của giai đoạn này có ảnh hưởng sâu sắc đến xử lý ngôn ngữ là mô hình các máy tự động (automaton) và các mô hình của lý thuyết thông tin hay xác suất: Máy tính điện tử bắt nguồn từ các mô hình máy Turing (1936), lý thuyết ngôn ngữ hình thức, mã hóa, entropy, …

Hai nhánh tách rời nhau từ cuối các năm 1950 đến đầu năm 1970: Hai kiểu xử lý khác biệt nhau rõ rệt: (a) kiểu hình thức (symbolic paradigm) cho văn bản như lý thuyết ngôn ngữ hình thức của Chomsky và trí tuệ nhân tạo, và (b) kiểu ngẫu nhiên (stochastic paradigm) cho tiếng nói như các phương pháp Bayes.

Bốn kiểu xử lý ngôn ngữ phổ biến trong thập kỷ 70 đến giữa thập kỷ 80:

(i) các *mô hình ngẫu nhiên* đóng một vai trò lớn, tiêu biểu là các mô hình Markov ẩn (HMM, Hidden Markov Model) trong xử lý tiếng nói;

(ii) *kiểu dựa trên logic* (logic-based paradigm);

(iii) *hiểu ngôn ngữ tự nhiên* (natural language understanding);

(iv) *mô hình hóa các cuộc đối thoại liên tục* (discourse modeling).

Cũng trong thời gian này xuất hiện các bài toán và tài nguyên chuẩn (Penn Treebank, WordNet, MUC, etc.).

Chủ nghĩa kinh nghiệm và các mô hình hữu hạn trạng thái từ giữa thập kỷ 80 đến giữa thập kỷ 90: Ở giai đoạn này đã có thể huấn luyện các mô hình hữu hạn trạng thái ra đời trong thập kỷ 60. Các mô hình xác suất và tiếp cận dựa vào dữ liệu (data-driven approach) xuất hiện trong hầu hết các nhiệm vụ của xử lý ngôn ngữ [Manning and Schutze, 1999], [Jelinek, 1998].

Xử lý văn bản và tiếng nói gặp nhau trong mười năm qua: Đây là giai đoạn tiến bộ vượt bậc với mô hình thống kê và tiếp cận dựa vào dữ liệu, với việc tăng trưởng nhanh chóng của tốc độ và bộ nhớ máy tính, của các ứng dụng dựa trên Web [Jurafsky and Martin, 2000]. Công nghệ xử lý tiếng nói không thể chỉ dựa trên các kỹ thuật xử lý tín hiệu, mà còn phải dựa cả trên việc hiểu ngôn ngữ. Do tham số của các mô hình thống kê hoặc các mô hình hữu hạn trạng thái đã có thể huấn luyện được từ các kho ngữ liệu lớn, nhiều mô hình kiểu này tiếp tục ra đời và được ứng dụng rộng rãi như Maximum Entropy (MaxEnt), Maximum Entropy Markov Model (MEMM), Conditional Random Fields (CRF), … [Cohen and McCallum, 2003]

Như đã trình bày sơ bộ ở trên, xử lí ngôn ngữ là một việc khó, phức tạp, chỉ có thể làm lâu dài theo nhiều bước tuần tự, chỉ có thể đạt được kết quả bước sau khi bước trước đã có kết quả. Chẳng hạn như các chương trình dịch tự động trên thế giới đã được theo đuổi hàng chục năm và chặng đường đến đích cuối vẫn còn rất xa. Nếu chúng ta muốn làm dịch tự động Anh-Việt, bắt buộc chúng ta đi qua các tầng của xử lý ngôn ngữ kể trên, và nói chung không thể hy vọng thời gian sẽ ngắn hơn nhiều so với người đi trước [Dorr et al., 2000].

Trên thế giới, nhiều tổ chức quốc tế, nhiều hiệp hội xử lí ngôn ngữ tự nhiên đã được thành lập với các hoạt động phong phú hàng năm, với lực lượng nghiên cứu đông đảo: ACL (Association Computational Linguistics), NAACL (North American Association on Computational Linguistics), EACL (Euro Association on Computational Linguistics), PACLIC (Pacific Association on Computational Linguistics), ICCL (International committee Computational Linguistics). Rất đáng chú ý là nhiều nguồn tài nguyên, ngữ liệu phong phú được tạo ra, được chia sẻ dù đôi khi dưới dạng thương mại nhưng với giá cả hợp lí, tiêu biểu là LDC (Linguistic Data Consortium, http://ww.ldc.upenn.edu).

Nhiều chính phủ đã đầu tư lớn cho xử lý ngôn ngữ trong CNTT (Mỹ, Nhật bản, Trung quốc, Singapore, etc.). Hãy thử nhìn đến các nước quanh Việt Nam. Ở Trung quốc, nghiên cứu về xử lý ngôn ngữ có từ lâu và hiện đang rất phát triển với sự đầu tư mạnh mẽ từ chính phủ. Họ đã làm ra nhiều công cụ, tài nguyên phong phú, như Wordnet cho tiếng Trung quốc, các hệ dịch giữa tiếng Trung quốc và qua tiếng một số nước, các ngân hàng ngữ liệu cho phân tích cú pháp như

Chinese Bank. Ngay với nước láng giềng Thailand, xử lý ngôn ngữ Thái đã là phần được đầu tư lớn của chính phủ tại National Electronics and Computer Technology Center (NECTEC), National research council of Thailand (NRCT), Thai Research Foundation (TRF) với những kết quả ban đầu về dịch máy, về POS tagging. Thailand đã có các phòng thí nghiệm lớn nghiên cứu về NLP, Thai Computational Linguistics Laboratories (http://www.tcllab.org).

Là người đi sau trong lĩnh vực xử lí ngôn ngữ, việc hiểu các công nghệ ngôn ngữ, các xu thế công nghệ, các nguồn ngữ liệu, kinh nghiệm và bài học từ các nước khác trong lĩnh vực này là hết sức quan trọng. *Biết, học và khai thác được chúng sẽ giúp ta cân nhắc chọn được đường đi hợp lí trong xử lý tiếng Việt*.

4. Tình hình và những vấn đề chính trong xử lý tiếng Việt

Hãy thử nhìn lại tình hình của chúng ta. Ngoài những việc đã làm và bước đầu làm được kể ở phần đầu, đã có những cố gắng trong nhiều nội dung khác nhau về xử lý ngôn ngữ và tiếng Việt. Trong các bài giảng về Trí tuệ Nhân tạo, về Lý thuyết Nhận dạng, về Xử lý Tín hiệu, về Khai phá Dữ liệu ở nhiều đại học, các nội dung và kỹ thuật xử lý ngôn ngữ đã ít nhiều được đề cập. Về xử lý tiếng nói và tiếng Việt, theo chúng tôi biết, hai tập thể làm nghiên cứu đã có những kết quả gần đây là Viện Công nghệ Thông tin và Trung tâm nghiên cứu quốc tế Thông tin đa phương tiện, truyền thông và ứng dụng (MICA) – Đại học Bách khoa Hà nội; một số kết quả ở một số trường Đại học là những đề tài tiến sĩ, thạc sĩ, mang tính chất tìm hiểu, chưa hệ thống và định hướng rõ ràng. Nghiên cứu về xử lý ngôn ngữ (văn bản) đã được theo đuổi bởi một số tập thể từ khá lâu (Đại học Bách khoa và Đại học Khoa học Tự nhiên thành phố Hồ Chí Minh, Đại học Bách Khoa Đà Nẵng, Đại học Bách khoa và Đại học Khoa học Tự nhiên Hà Nội, Trường Đại học Công nghệ, Viện Ứng dụng Công nghệ, Viện Công nghệ Thông tin, công ty Lạc Việt,…) về các vấn đề dịch máy, các bài toán cơ bản của xử lý tiếng Việt [Dien et al., 2001; Dien, 2003], [Huyen et al., 2003], và gần đây là tóm tắt văn bản [Minh, 2004; Minh et al., 2004], [Huong, 2004], tìm kiếm và trích chọn thông tin [Bao and Funakoshi, 1998], phân loại và chia nhóm văn bản [Bao and Binh, 2001], khai phá Web [Hieu, 2005], gióng hàng văn bản [Huyen et al., 2003], mô hình từ điển điện tử [Bao et al., 2003], xây dựng kho ngữ liệu [Dien, 2002], … và gần đây nhất là đề tài nhà nước "Nghiên cứu phát triển công nghệ nhận dạng, tổng hợp và xử lý ngôn ngữ tiếng Việt" giai đoạn 2001-2004 trong chương trình quốc gia KC-01 [Khang, 2004].

Bên ngoài Việt Nam, cũng có những nỗ lực về xử lý tiếng Việt, như nhóm dịch Anh-Việt của tiến sỹ Phạm Hải và bè bạn (Mỹ) khởi đầu từ đầu các năm 1990, của tiến sỹ Lê Tăng Hồ và phần mềm tổng hợp tiếng Việt VVV (Canada), … đặc biệt là của các cán bộ và nghiên cứu sinh Việt Nam tại Viện Khoa học và Công nghệ Tiên tiến Nhật bản (JAIST) với 6 nghiên cứu sinh về xử lý ngôn ngữ, hợp tác theo một kế hoạch thống nhất. Số nghiên cứu sinh về xử lý ngôn ngữ mới tốt

nghiệp như Lê Thanh Hương (Anh) [Huong, 2004], Nguyễn Thị Minh Huyền (Pháp), Hồ Bảo Quốc (Pháp), Nguyễn Lê Minh (Nhật) [Minh, 2004], và số sắp tốt nghiệp trong 1-2 năm tới đây ước tính khoảng 10 người.

Ngoài những kết quả ban đầu, sau đây có thể là một vài đặc điểm chính về hoạt động xử lý ngôn ngữ của chúng ta:

- Thường tập trung vào làm các sản phẩm cho người dùng cuối với nhiều kỳ vọng vào các sản phẩm dịch máy, một loại sản phẩm khó làm và cần làm dài hạn với những phương pháp hợp lý.
- Ít các nghiên cứu nền tảng, thiếu phát triển "hạ tầng cơ sở" cho xử lý ngôn ngữ như công cụ và tài nguyên: từ điển (dùng cho máy), kho ngữ liệu, …, những thứ đã được cả cộng đồng quốc tế xác định là không thể thiếu trong xử lý ngôn ngữ.
- Phần đông là các nghiên cứu ngắn hạn và đơn lẻ ở mức đề tài thạc sỹ, tiến sỹ với nhiều hạn chế về thời hạn và điều kiện. Đa số mới xây dựng được mô hình, thử và kiểm tra trên những tập ngữ liệu nhỏ. Những kết quả đạt được ở đây còn xa với mức sử dụng được trong thực tế.
- Rất có thể nhiều nhóm đã bắt đầu với sự khảo sát chưa đầy đủ, hoặc tiến hành công việc khi có thể còn thiếu kiến thức. Do vấn đề mới và phức tạp, các hội đồng đánh gía thẩm định các đề tài về xử lý ngôn ngữ còn chưa có chuyên gia, chưa thật rõ hết cái có thể và cái chưa thể làm được, người làm đi đường ngắn hay đường vòng, …
- Đáng băn khoăn hơn cả là các nỗ lực của chúng ta chưa được liên kết, thiếu chia sẻ, phân công, hợp tác theo một lộ trình có kế hoạch, thiếu "kim chỉ nam" về xử lý tiếng Việt, và không có tính kế thừa về kết quả giữa các tầng của xử lý ngôn ngữ tự nhiên. Nếu hình dung công việc trong các tầng của xử lý ngôn ngữ được đánh số từ A đến Z, thì hầu hết các việc làm ra cho người dùng cuối đều ở quãng từ R, S, … trở đi, mà muốn làm mấy việc này thì đều cần kết quả của tất cả các bước từ A đến tận P, Q. Hiềm nỗi mỗi việc từ A, B, … đến P, Q muốn làm tốt đều đòi hỏi một nhóm người làm trong một vài năm. Vì vậy, nếu ai cũng phải làm từ A đến gì đấy tận P, Q, có lẽ sẽ không ai có thể làm ra các sản phẩm R, S, …, Z đủ tốt. Dù bây giờ hay năm, mười, hai mươi năm sau cũng vậy.

5. Kết luận

Chúng ta ai cũng biết xử lý tiếng Việt chỉ có thể do người Việt làm, không thể mua được từ bên ngoài. Ngoài ra, xử lý tiếng Việt là công việc phải làm trên đường dài gồm nhiều chặng ngắn với các đích chọn lựa kỹ lưỡng, và cần được nhà nước hỗ trợ. Trước mắt, trong kế hoạch 2006-2010, hai mục tiêu của dự án đầu là:

1. Xây dựng và phát triển một số sản phẩm tiêu biểu về xử lý tiếng Việt và tìm kiếm thông tin trên Internet bằng tiếng Việt cho đông đảo người sử dụng máy tính và Internet.
2. Xây dựng các công cụ và nguồn tài nguyên thiết yếu, với vai trò hạ tầng cơ sở, để thực hiện mục tiêu 1 của dự án này và các phát triển lâu dài của công nghệ thông tin của nước nhà.

Mặc dù việc phải tạo các sản phẩm cho người dùng cuối là cấp bách và đích cuối cùng ta cần đến, chúng ta vẫn luôn phải chú ý đầu tư cho việc tạo ra các công cụ và tài nguyên cho xử lý tiếng Việt, vì chỉ có như vậy ta mới có thể làm được các sản phẩm có thể dùng được trong những năm về sau.

Lời cám ơn

Xin chân thành cám ơn các ý kiến thảo luận, góp ý cho bản thảo tài liệu này của các đồng nghiệp: Ngô Trung Việt, Phạm Ngọc Khôi, Nguyễn Lê Minh, Phan Xuân Hiếu, Lê Anh Cường, Lê Minh Hoàng, Đỗ Bá Phước, Ngô Thanh Nhàn, Trần Hữu Dũng, Nguyễn Hoàng, Hồ Văn Tiến.

Tài liệu tham khảo 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 Natural Language Processing, *Vietnam-Japan Symposium on Fuzzy Systems and Applications,* NCST, Hanoi, 7-9 December, 297-304.

Bao, H.T., Funakoshi, K. (1998). Information Retrieval Using Rough Sets, *Journal of Japanese Society for Artificial Intelligence*, JSAI, Vol. 13, N. 3, 424-433

Bao, H.T., Binh, N.N. (2002). Nonhierarchical Document Clustering by a Tolerance Rough Set Model, *International Journal of Intelligent Systems*, John Wiley & Sons, Vol. 17, No. 2, 199-212.

Bao, H.T., Tuan, N.A., Son, N.C. (2003). Issues in Construction of a Vietnamese Machine Tractable Dictionary, *First National Symposium on Research, Development, and Applications of Information and Communication Technology*, 253-263.

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 Conference on Knowledge Discovery and Data Mining 2003, KDD-03, Washington D.C., Aug. 2003.

Cole, R., Mariani, J., Uszkoreit, H., Varile, G., Zaenen, A., Zampolli, A., Zue, V. (1997). *Language Technology – A Survey of the State of the Art*.

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 Natural Language Processing Pacific Rim Symposium* (NLPRS2001), Tokyo (Japan, 27-30 November 2001, p. 749-756.

Dien, D. (2002). Building a Training Corpus for Word Sense Disambiguation in English-to-Vietnamese Machine Translation*, Workshop Machine Translation in Asia at COLING 2002*.

Dien, D., Kiem, H. (2003). POS-Tagger for English-Vietnamese Bilingual Corpus, *Proceedings of HLT-NAACL 2003* (Human Language Technology- North American Chapter of the Association for Computational Linguistics).

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 Language Processing, Computational Linguistics and Speech Recognition*, Prentice Hall.

Hieu, P.X., Horiguchi, S., Bao, H.T. (2005). Conditional Models for Automatic Data Integration from the Web, *International Journal on Business Intelligence and Data Mining (in press).*

Huong, L.T. (2004). *Investigation into an Approach to Automatic Text Summarisation*, Ph.D. dissertation, Middlesex University, U.K.

Huyen, N.T.M., Laurent Romary, Luong, V.X. (2003). A Case Study in POS Tagging of Vietnamese Texts, *Annual Conference on Natural Language Processing TALN 2003*, Batz-sur-Mer, 11-14 June, 2003.

Khang, B.H. et al. (2004). *Báo cáo Tổng kết Khoa học và Kỹ thuật Đề tài Nghiên cứu Phát triển Công nghệ Nhận dạng, Tổng hợp và Xử lý Ngôn ngữ Tự nhiên*. Chương trình KC-01.

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 Hidden Markov Model, *ACM Transactions on Asian Language Information Processing,* Vol. 3, Issue 3, 146-158.

Minh, N.L., Shimazu, A., Horiguchi, S., Bao, H.T., Fukushi, M. (2004). Probabilistic Sentence Reduction Using Support Vector Machines, *The 20th International Conference on Computational Linguistics COLING 2004*, 23-27 August, Geneva, 743-749.

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

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.