

Tiếp theo chương trình nghiên cứu-phát triển về Xử lý Tiếng nói và Văn bản tiếng Việt (2007-2009)

Lương Chi Mai, Hồ Tú Bảo
Viện CNTT, Viện KHCN VN
Đại diện Đề tài VLSP

Nội dung trình bày

- Nhìn lại các kết quả của VLSP (2007-2009)
 - Mục tiêu
 - Nội dung đã và chưa làm được
- Đề xuất
 - Tiếp tục từ phía VLSP
 - Những nội dung, sản phẩm cần có
- Chính sách từ Nhà nước, Doanh nghiệp...

Nhìn lại đề tài VLSP

- | | |
|--|--|
| <ul style="list-style-type: none"> ▪ Xử lý NN TN&VB phải theo các bước tuần tự, liên quan chặt chẽ ▪ Xử lý tiếng Việt <ul style="list-style-type: none"> □ Thiếu lộ trình chung □ Các nhóm làm việc biệt lập, thiếu hợp tác và kế thừa □ Thiếu tài nguyên và công cụ cho VLSP <ul style="list-style-type: none"> – Từ điển – Kho ngữ liệu (annotated corpora) – Các công cụ xử lý cơ bản ▪ Chỉ làm được trong những chương trình dài hạn, được tổ chức có tính kế thừa. ▪ Hợp tác, phát triển từng bước, chia sẻ của nhiều tập thể và cá nhân. | <ul style="list-style-type: none"> ▪ Nội dung, phương tiện, tài nguyên cơ bản là gì? ▪ Lộ trình thế nào? Kế hoạch ngắn hạn, dài hạn? ▪ Đây là những nghiên cứu nền tảng phải làm cho tiếng Việt? Đây là những ứng dụng tiêu biểu nên làm và cần làm? ▪ Cần học gì, mua gì từ bên ngoài? Và cần phải tự làm những gì? ▪ Tổ chức, hợp tác, chia sẻ thế nào? ▪ Xây dựng một Chương trình, Dự án nhà nước? ▪ Thời điểm đó (2005) chưa nhiều sản phẩm về XLNN (Lạc Việt, EVTRAN) |
|--|--|

3

Hội thảo XLNN lần 1, 29/3/2005

Nhìn lại các kết quả của VLSP: Mục tiêu lâu dài



1. Nghiên cứu và xây dựng một số sản phẩm tiêu biểu về xử lý tiếng nói và văn bản tiếng Việt, nhằm cho đồng đảo người sử dụng máy tính và Internet.
2. Xây dựng các tài nguyên và công cụ thiết yếu nhất cho xử lý tiếng nói và văn bản tiếng Việt, nhằm thực hiện mục tiêu 1 và dùng cho nghiên cứu-phát triển lâu dài.

5

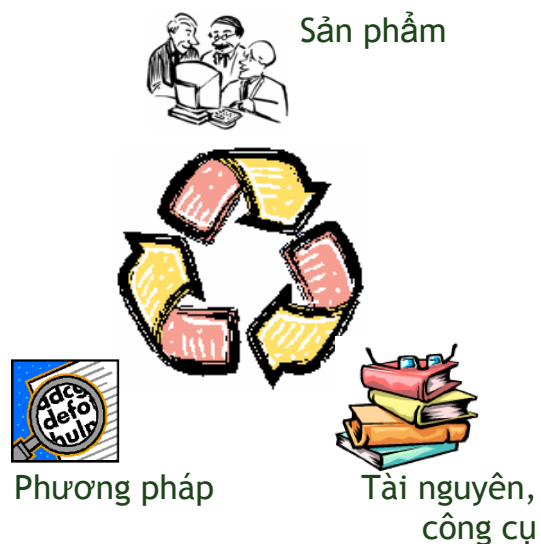
Nhìn lại các kết quả của VLSP



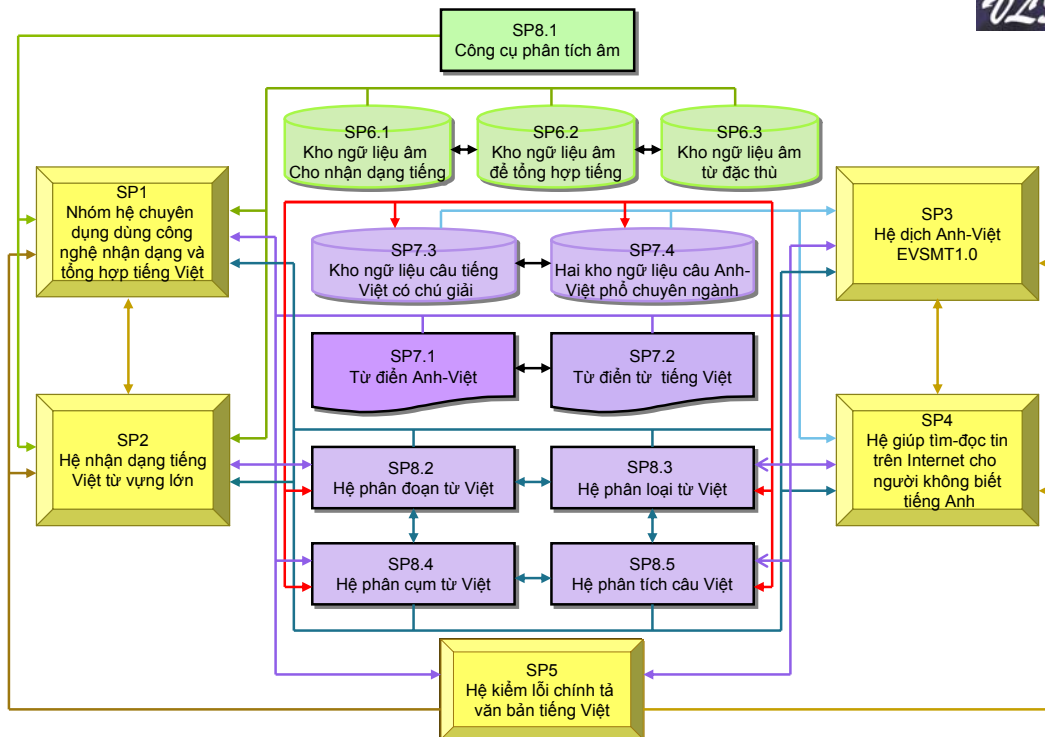
Xây dựng một cấu trúc nền tảng, đặc biệt các hướng dẫn (guideline), phương pháp và tài nguyên công cụ cho cộng đồng nghiên cứu về VLSP

Phát triển một số sản phẩm tiêu biểu định hướng cho người dùng cuối

Lần đầu tiên có một đề tài về VLSP gồm 8 đơn vị có kinh nghiệm và làm việc tích cực

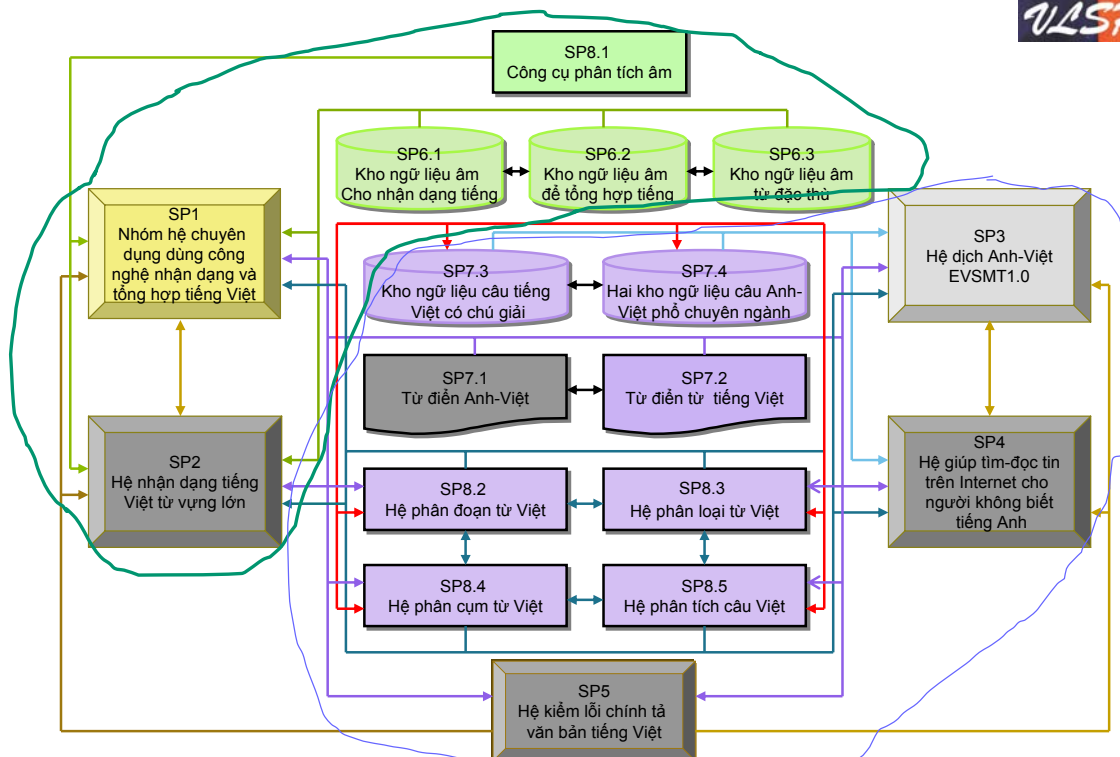


6



Sản phẩm thiết kế cho giai đoạn 5 năm 2006-2010

7

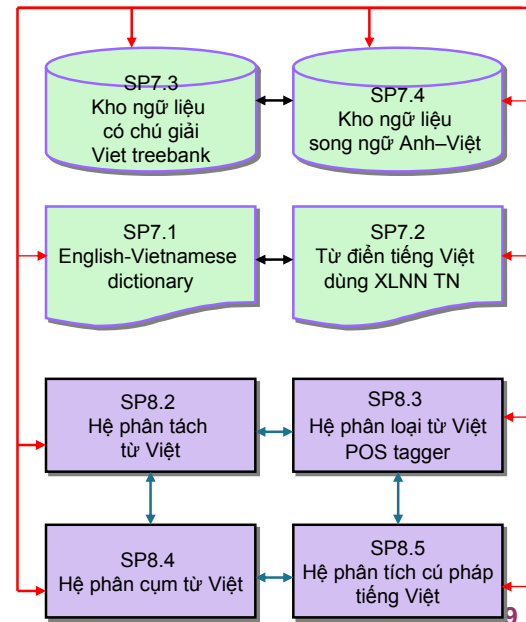


Sản phẩm thực hiện cho giai đoạn 2 năm 2007-2009

8

Nhìn lại các kết quả của VLSP (tiếp)

- ❑ Tạo một cộng đồng có trình độ về nghiên cứu và phát triển (8 đơn vị, 2 miền)
- ❑ Các vấn đề giải quyết: nền tảng (công nghệ lõi, công cụ, tài nguyên).
 - VD: Dựn "chuẩn" → Thách thức: "Chuẩn" trong sự phát triển bền vững
 - Sự cần thiết có chuẩn cho VLSP: tham khảo các quan điểm khác nhau về tiếng Việt, phân tích và lựa chọn một quan điểm thích hợp
 - Các công cụ đều được xây dựng dựa trên cùng một quan niệm về từ, gắn nhãn, câu và trên cùng tài nguyên
 - Cùng sử dụng những phương pháp thống kê trong Học máy để xây dựng công cụ và tài nguyên kho ngữ liệu
- ❑ Theo sát các xu hướng phát triển công nghệ NLP trên thế giới
- ❑ Công cụ và tài nguyên nhằm chia sẻ cho cộng đồng
- ❑ Có thể dựa trên các tài nguyên, công cụ, công nghệ lõi đã có để phát triển tiếp



VLSP website khối văn bản

<http://vlsp.vietlp.org:8080/demo/>

ĐỀ TÀI VLSP
Nhánh đề tài XỬ LÝ VĂN BẢN

Trang chủ | Từ điển tiếng Việt | Phân đoạn → Phân loại → Phân cụm | **Phân tích cú pháp** | Giới thiệu

Phân tích cú pháp câu

Hệ thống tiến hành phân tích cú pháp một câu tiếng Việt theo hai phương pháp: "Phương pháp tiếp cận dựa trên cấu trúc tâm (HPSG)" và "Phương pháp tiếp cận dựa trên học máy thống kê".
Xin mời nhập vào một câu tiếng Việt:

Hà Nội tăng trưởng thấp nhất trong các thành phố trung ương.

☒ Phương pháp tiếp cận dựa trên cấu trúc cụm từ hướng trung tâm (HPSG)
☐ Phương pháp tiếp cận dựa trên học máy thống kê

Phân tích...

Kết quả phân tích

Câu bạn vừa nhập: Hà Nội tăng trưởng thấp nhất trong các thành phố trung ương.
Có 2 cách phân tích:

1. (S (VP (V hà) (N Nội)) (VP (V tăng trưởng) (AP (A thấp) (R nhất)) (PP (C trong) (NP (N các) (N thành phố) (N trung ương))
2. (S (V hà) (NP (N Nội)) (VP (V tăng trưởng) (AP (A thấp) (R nhất)) (PP (C trong) (NP (N các) (N thành phố) (N trung ương))

Mudim v0.8 ☐ Tất ☐ VNI ☐ Telex ☐ Vqr ☒ Tổng hợp ☐ Tự động ☒ Chính tả ☒ Bỏ dấu kiểu mới [☐ Bật/Tắt (F9) ☐ Ẩn/Hiện (F8)]

Nhìn lại các kết quả của VLSP (tiếp)

- Tài nguyên tiếng nói
 - Nguyên tắc xây dựng CSDL tiếng nói (sau một thời gian mới có thể xác định được tiêu chuẩn này)
 - Công cụ trợ giúp, quản lý thu âm
 - Một số CSDL âm cho tổng hợp và nhận dạng tiếng Việt
- Một số engines
 - TTS tiếng Việt (HMM, concatenation) → định hướng phát triển
 - Phương pháp – huấn luyện và decode → định hướng phát triển
 - Sẵn sàng tích hợp trong nhiều môi trường ứng dụng khác nhau
- Có thể tham gia hợp tác với cộng đồng quốc tế

11

Nhìn lại các kết quả của VLSP (tiếp)

- Các minh chứng)
 - Segmenter:
 - Postarger
 - VieTreebank

12

Nhìn lại các kết quả của VLSP:

Tiếp tục thực hiện

- VLSP đang tiếp tục công bố tài nguyên (VieTreebank, tiếng nói VOV) và các công cụ cho cộng đồng nghiên cứu
 - LDC (đã dịch tài liệu sang tiếng Anh, các agreement form)
 - Hiện tại cho lấy qua trang của VLSP
<http://vlsp.vietlp.org:8080/demo/?page=resources&lang=vi>
<http://www.loria.fr/~lehong/tools/vnToolkit.php>
- Đang implement phần TTS tiếng Việt trên Web cho cộng đồng tích hợp vào ứng dụng
- Phát triển để chia sẻ thông tin trong giới nghiên cứu và Đại học
 - Thông tin các luận án, luận văn về XLNNTV
 - Thông tin cập nhật kiến thức advanced trong XLNN: văn bản và tiếng nói
 - Các công cụ được cập nhật (tách từ, gán nhãn, gộp nhóm từ loại,...)

13

Nhìn lại các kết quả của VLSP:

Một số nhận xét

- Khối lượng công việc quá nặng ở một số đầu mục liên quan đến xây dựng các kho ngữ liệu (do thiết kế cho Dự án 5 năm)
 - Minh chứng về số lượng của các kho ngữ liệu văn bản
 - Một số ví dụ minh họa về đầu tư và độ lớn của các kho ngữ liệu tiếng nói, chứng minh tính không hợp lý
 - kinh phí không đủ để vừa làm donation lại vừa phải phát triển sản phẩm mang tính cạnh tranh
- Yêu cầu sản phẩm đầu cuối là không thực tế (kinh nghiệm của các nước tiên tiến) từ phía các Đề tài KHCN – nên đặt vấn đề về phát triển công nghệ lõi → chuyển giao công nghệ cho doanh nghiệp và phát triển thành sản phẩm
- Minh họa về các công nghệ lõi được phát triển từ Đề tài
 - TTS: engine có thể được tích hợp trong nhiều ứng dụng → qua Bộ TTTT tích hợp cho người mù
 - Vẫn cần được đầu tư nâng cấp (tiếng nói mới, độ tự nhiên...) → tiếp tục chuyển giao công nghệ
 - Trách nhiệm của Bộ KHCN: đầu mối giới thiệu cho doanh nghiệp và ủng hộ các dự án SXTN
 - Vấn đề IP!

14

Một số ý kiến về lộ trình tiếp theo

- Một số ví dụ minh họa về đầu tư và độ lớn của các kho ngữ liệu tiếng nói cho cộng đồng, chỉ ra tính không hợp lý
 - kinh phí không đủ để vừa làm donation lại vừa phải phát triển sản phẩm mang tính cạnh tranh
- Minh họa về các công nghệ lõi được phát triển từ Đề tài
 - TTS: engine có thể được tích hợp trong nhiều ứng dụng → qua Bộ TTTT tích hợp cho người mù
 - Vẫn phải phát triển tiếp → đã được đầu tư rồi sao vẫn đăng ký tiếp – tiếng nói mới, độ tự nhiên
 - Trách nhiệm của Bộ KH-CN: đầu mỗi giới thiệu cho doanh nghiệp và ủng hộ các dự án SXTN
 - Vấn đề IP

15

Một số ý kiến về lộ trình tiếp theo

- XLNNTV là bước đi đường dài, cần chuẩn bị cơ bản
- Những công việc và đầu tư tiếp theo: phải dựa trên những cái đã có → xác định phải làm tiếp những vấn đề gì và phát triển tiếp

- Giới Nghiên cứu và Đại học → học thuật, cung cấp kiến thức, cung cấp con người, công nghệ lõi

←

- Đầu tư từ Nhà nước

- Giới doanh nghiệp → sản phẩm

- Đầu tư từ Doanh nghiệp

16

Một số ý kiến về lộ trình tiếp theo

- Quản lý và phát triển tài nguyên
 - Tổ chức non-profit → cần có tài trợ của Nhà nước, doanh nghiệp...
 - Non-profit: để quản lý, điều hành, phân phối kho ngữ liệu (cho mục đích khác nhau: nghiên cứu, thương mại hóa)
 - Ví dụ về LDC
 - Ví dụ về Nat. Project của Đại học Kyoto, Nhật bản
 - Ví dụ của Academia Sinica
 - Hiện tại: <http://vlsp.vietlp.org:8080/demo/?page=home>
- Nghiên cứu, cách tổ chức phát triển các kho ngữ liệu
 - Các tiêu chí đánh giá các kho ngữ liệu, kể cả độ lớn → tư vấn cho Bộ KHCN về đầu tư kinh phí
 - Phân cấp cho các Cty làm dữ liệu
 - Phát triển các kho ngữ liệu có sự tham gia của cộng đồng – cơ chế kiểm soát, cập nhật
 - Quản lý kinh phí: kinh phí mua dữ liệu từ các cơ sở
- Nơi thông báo ai làm gì, thi đua lành mạnh (nếu làm cùng vấn đề) hoặc tránh chồng chéo...

17

Một số ý kiến về lộ trình tiếp theo

- 

Sản phẩm





**Công cụ,
dữ liệu,
tài nguyên,
phương tiện**



**Công nghệ
Phương pháp**
- Tiếng nói
 - Giao tiếp tiếng nói tự nhiên
 - Phát triển các kho ngữ liệu đặc thù
 - Nhận dạng: Ứng dụng từng bước trong các bài toán hạn chế
 - Sản phẩm có thị trường như dịch tiếng nói trong giao tiếp du lịch, hỏi thông tin
 - Tìm kiếm bằng hỏi đáp tiếng nói (tích hợp với các hệ thống Q&A), tìm nhạc (Zing)
 - Ứng dụng nhúng (chip)
 - Tổng hợp: rất nhiều ứng dụng và sẵn sàng nhận các đặt hàng
 - Làm chủ công nghệ và có khả năng triển khai dễ dàng sang các ứng dụng khác nhau
 - Có cơ chế an toàn trong chuyển giao công nghệ

18

Một số ý kiến về lộ trình tiếp theo



- Tiếp tục nâng cấp hoàn thiện các công cụ: tách từ, gán nhãn từ loại, gộp nhóm từ, phân tích cú pháp
- Văn bản
 - Text mining
 - Topic modeling
 - Cross Language Information Retrieval
 - Q&A System
 - Semantic Content Based Video Recommendation (SCVR)

19

Quan sát

- Kiểm lỗi chính tả (KLCT)
- Tìm kiếm
- Phân loại văn bản
- Text mining
 - Nhóm ĐHCN ĐHQGHN, JAIST
- Web mining, semantic Web
 - ĐHBK, ĐHKHTN TpHCM,
- Dịch máy (MT): khó khăn của làm MT?
 - Lạc Việt
 - EVTRAN
 - BOCOCHAN
 - ...

20