TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN
KHOA CÔNG NGHỆ THÔNG TIN
LỚP CỬ NHÂN TÀI NĂNG

HOÀNG TRỌNG NGHĨA – 0512031

KẾT HỢP CÁC NGUỒN TÀI NGUYÊN KHÁC NHAU ĐỂ XỬ LÝ NHẬP NHẰNG NGỮ NGHĨA TRONG DỊCH MÁY ANH – VIỆT

KHÓA LUẬN TỐT NGHIỆP CỬ NHÂN CNTT

GIÁO VIÊN HƯỚNG DẪN

PGS. TS. ĐINH ĐIỀN

KHÓA 2005 – 2009

NHẬN XÉT CỦA GIÁO VIÊN HƯỚNG DẪN

.....
.....
.....
.....
.....
.....
.....
.....
.....
.....
.....
.....
.....
.....
.....
.....
.....
.....
.....
.....
.....
.....
.....
.....
.....
.....
.....
.....
.....
.....
.....
.....
.....
.....
.....
.....
.....
.....

NHẬN XÉT CỦA GIÁO VIÊN PHẢN BIỆN

.....
.....
.....
.....
.....
.....
.....
.....
.....
.....
.....
.....
.....
.....
.....
.....
.....
.....
.....
.....
.....
.....
.....
.....
.....
.....
.....
.....
.....
.....
.....
.....
.....
.....
.....
.....

Khóa luận đáp ứng yêu cầu của Khóa luận cử nhân CNTT.
TP. HCM, ngày … tháng … năm 2009
Giáo viên phản biện
[Ký tên và ghi rõ họ tên]

LỜI CẢM ƠN

Để hoàn thành luận văn này, ngoài những nỗ lực, cố gắng của bản thân, em còn nhận được rất nhiều sự quan tâm giúp đỡ, hỗ trợ tận tình và sự động viên từ các thầy cô, bạn bè và gia đình.

Trước tiên, em xin chân thành cảm ơn PGS. TS. Đinh Điền là người đã tận tình chỉ dẫn, động viên và gợi mở cho em nhiều vấn đề liên quan đến đề tài luận văn của em.

Em cũng xin chân thành cảm ơn các thầy cô trong Khoa CNTT đã truyền đạt cho em những tri thức và kinh nghiệm quý báu trong trên giảng đường Đại học và trong quá trình làm luận văn.

Con xin chân thành cảm ơn bố mẹ đã tạo mọi điều kiện tốt nhất cho con học tập và động viên, khích lệ con trong suốt quá trình thực hiện luận văn.

Và cuối cùng, tôi xin gửi lời cảm ơn đến tất cả những bạn bè, những người đã hỗ trợ và giúp tôi hoàn thành luận văn này.

TP. Hồ Chí Minh, tháng 7 năm 2009

Hoàng Trọng Nghĩa

LỜI NÓI ĐẦU

Một trong những trở ngại lớn nhất của dịch máy tự động chính là việc phân tích nhập nhằng ngữ nghĩa trong ngôn ngữ. Trong hầu hết các hệ xử lý ngôn ngữ, việc phân tích ngữ nghĩa (semantic analyzer) hay còn gọi là "khử nhập nhằng ngữ nghĩa của từ" (Word Sense Disambiguation / Word Sense Discrimination, viết tắt là WSD) là bài toán khó khăn nhất và cũng là bài toán trọng tâm mà đến này thế giới vẫn chưa giải quyết ổn thỏa được.

Hiện nay, trên thế giới đã có rất nhiều mô hình với các cách tiếp cận khác nhau, chủ yếu gồm các hướng sau :

- Dựa trên trí tuệ nhân tạo (AI-based) : Đây là cách tiếp cận sớm nhất (1960) dựa trên cơ sở lý thuyết về mạng ngữ nghĩa, khung ngữ nghĩa và các ý niệm nguyên thủy. Hầu hết các trí thức về ngữ nghĩa trong cách tiếp cận này đều được xây dựng bằng tay (hand-coded rules) dẫn đến việc hạn chế khi mô phỏng thế giới thực.

- Dựa trên cơ sở trí thức (Knowledge-based) : Để khắc phục tình trạng thiếu tri thức như trong cách tiệp cận dựa trên luật, vào đầu thập niên 1980, người ta đã chuyển sang khai thác tri thức tự động từ các từ điển điện tử (MRD : Machine Readable Dictionaries) như : từ điển đồng nghĩa (thesaurus), WORDNET, LDOCE, LLOCE, … Tuy nhiên, các cơ sở trí thức nói trên cũng chỉ là những nguồn thông tin đóng vài trò tham khảo và thường có độ bao phủ thấp trên ngôn ngữ.

- Dựa trên ngữ liệu (Corpus-based) : Hướng tiếp cận này chủ yếu rút ra những quy luật xử lý ngữ nghĩa (bằng thống kê, bằng máy học, ...) từ những kho ngữ liệu lớn (tập hợp các văn bản được gán nhán ngữ nghĩa chính xác) đã có sẵn và

áp dụng những luật này cho các trường hợp mới. Với cách tiếp cận này, các nguồn tri thức phục vụ cho việc xử lý nhập nhằng ngữ nghĩa sẽ được chọn lựa và sử dụng hiệu quả nhất. Trở ngại của hướng tiếp cận này là đòi hỏi ngữ liệu lớn mà đã được gán nhãn ngữ nghĩa trước, là công việc đòi hỏi chi phí cao và tiêu tốn rất nhiều thời gian, sức lực.

Gần đây, cách tiếp cận dựa trên nhiều nguồn tài nguyên, ngữ liệu kết hợp với trí thức có sẵn đang được nhiều nhà ngôn ngữ học – máy tính quan tâm. Với mục đích làm rõ tính khả thi của cách tiếp cận đó, luận văn này hướng đến một mô hình xử lý nhập nhằng ngữ nghĩa dựa trên việc kết hợp các nguồn tài nguyên : văn bản song ngữ, văn bản đơn ngữ và hệ thống nhãn ngữ nghĩa LLOCE. Theo đó, hệ thống xử lý nhập nhằng ngữ nghĩa sẽ được huấn luyện dựa trên ngữ liệu có gán nhãn thu thập tự động bằng cơ chế liên kết từ trên văn bản song ngữ (có kết hợp tham khảo trí thức ngữ nghĩa của LLOCE) và sau đó, hệ thống được nâng cấp chất lượng dần dần dựa trên việc khai thác văn bản đơn ngữ.

Luận văn được tổ chức thành 5 chương với nội dung như sau :

- Chương 1 : TỔNG QUAN – Giới thiệu về bài toán xử lý nhập nhằng ngữ nghĩa và các phương pháp phổ biến.
- Chương 2 : CƠ SỞ LÝ THUYẾT – Giới thiệu các cơ sở lý thuyết ngôn ngữ, tin học cần sử dụng.
- Chương 3 : MÔ HÌNH XỬ LÝ NHẬP NHẰNG NGỮ NGHĨA – Giới thiệu các cách tiếp cận trước đây, mô tả phương pháp đề xuất.
- Chương 4 : KẾT QUẢ THỬ NGHIỆM – Báo cáo và trình bày các kết quả thử nghiệm của phương pháp đề xuất
- Chương 5 : KẾT LUẬN VÀ HƯỚNG MỞ RỘNG – Nêu lên nhận xét, đánh giá chung cho phương pháp đề xuất và đưa ra hướng mở rộng.

MỤC LỤC

CHƯƠNG 1 – TỔNG QUAN

1.1 GIỚI THIỆU BÀI TOÁN XỬ LÝ NHẬP NHẰNG NGỮ NGHĨA

1.1.1 PHÁT BIỂU BÀI TOÁN

Như chúng ta đã biết, nhu cầu khử nhập nhằng ngữ nghĩa xuất phát từ thực tế là trong hầu hết các ngôn ngữ, có rất nhiều từ sẽ mang những sắc thái ý nghĩa khác nhau nếu đặt trong các ngữ cảnh khác nhau. Chính vì thế, vấn đề xử lý nhập nhằng ngữ nghĩa được đặt ra với mục đích tìm kiếm một cơ chế tự động có khả năng hiểu được ngữ cảnh xung quanh một từ đa nghĩa và chỉ ra được ý nghĩa tương thích nhất của nó.

Để dễ hình dung, chúng ta xét ví dụ sau :

- His mansion is located somewhere near the bank of a river.
- She is going to the bank to draw some money.

Ở ví dụ trên, ta thấy trong cả hai câu từ *bank* đều được dùng dưới hình thức danh từ nhưng lại mang hai sắc thái ý nghĩa khác hẳn nhau. Trong câu thứ nhất, từ *bank* có ý nghĩa là bờ sông, chúng ta biết được điều này là nhờ vào ngữ cảnh xung quanh nó có từ *river*. Ở câu thứ hai, từ *bank* lại mang nghĩa ngân hàng do ngữ cảnh của nó có từ *money*.

1.1.2 TẦM QUAN TRỌNG

Như vậy, rõ ràng là để hiểu ý nghĩa của một từ thì ta cần phải phân tích ngữ cảnh của nó. Vấn đề ở đây là làm sao để lập trình cho máy tính hiểu được ngữ cảnh đó. Đây chính là một trong những vấn đề trọng tâm của xử lý ngôn ngữ tự nhiên. Ngay

từ buổi ban đầu của lĩnh vực này, xử lý nhập nhằng ngữ nghĩa đã thu hút mạnh mẽ sự quan tâm và nghiên cứu của giới chuyên môn. Về bản chất, xử lý nhập nhằng ngữ nghĩa chỉ là một bài toán trung chuyển chứ không thật sự là nền tảng của một ứng dụng xử lý ngôn ngữ tự nhiên cụ thể nào. Tuy nhiên, đó lại là một bước thiết yếu không thể bỏ qua trong hầu hết các ứng dụng liên quan đến ngôn ngữ tự nhiên :

- Dịch máy tự động : Khử nhập nhằng ngữ nghĩa là một vấn đề thiết yếu của những hệ thống dịch tự động do các từ đa nghĩa trong ngôn ngữ đích có thể được dịch sang nhiều cách khác nhau ở ngôn ngữ nguồn. Ví dụ : Khi dịch từ tiếng Anh sang tiếng Việt, từ "*plant"* có thể được dịch là "*nhà máy*" hay "*thực vật*" tùy vào ngữ cảnh tương ứng.

- Rút trích thông tin tự động : Giả thiết chúng ta cần thu thập các văn bản liên quan đến một vấn đề cụ thể nào đó, nếu những từ khóa mà ta dùng để tìm kiếm lại mang tính đa nghĩa thì rõ ràng cần phải có sự khử nhập nhằng ngữ nghĩa để hạn chế rút trích ra những thông tin không liên quan đến chủ đề ta quan tâm. Ví dụ : Khi ta sử dụng từ khóa "*arms, weapon*" để tìm kiếm thì hệ thống cần phải hiểu được rằng "*arms*" ở đây mang sắc thái ý nghĩa "*vũ trang, vũ khí*" chứ không phải là một bộ phận của cơ thể - "*cánh tay*".

- Xử lý văn bản tự động : Khi hệ thống tiến hành kiểm lỗi chính tả, ngữ pháp thì khử nhập nhằng ngữ nghĩa cũng đóng vai trò quan trọng. Ví dụ : Khi hệ thống chuẩn hóa văn bản thô, "*HE READS THE TIMES*", từ *"TIMES"* cần được hiểu là "*thời báo*" để có thể được chuẩn hóa thành "*He reads the Times*" chứ không phải "*He reads the times*".

- Tổng hợp và nhận dạng tiếng nói : Đối với tổng hợp tiếng nói thì khử nhập nhằng ngữ nghĩa là cần thiết để máy tính có thể phát âm chính xác. Đối với nhận

dạng tiếng nói thì khử nhập nhằng ngữ nghĩa cũng cần thiết để phân đoạn từ và phân biệt từ đồng âm khác nghĩa.

1.2 LỊCH SỬ QUÁ TRÌNH NGHIÊN CỨU

Vào những ngày đầu của xử lý ngôn ngữ tự nhiên (1950), xử lý nhập nhằng ngữ nghĩa là một trong những bài toán đầu tiên được quan tâm và thu hút rất nhiều công trình nghiên cứu từ giới chuyên môn. Vào khoảng thập niên 60, trong giới chuyên môn lại có ý kiến cho rằng việc khử nhập nhằng ngữ nghĩa là không khả thi; điển hình là báo cáo của Bar – Hiller (ALPAC, 1996). Trong báo cáo này, Bar – Hiller chỉ ra những trường hợp không thể xác định ngữ nghĩa của từ đa nghĩa một cách tự động với những hướng tiếp cận lúc bấy giờ. Kết quả của báo cáo đó là hầu hết các nghiên cứu liên quan đến lĩnh vực xử lý ngôn ngữ tự nhiên đều bị bỏ dở.

Mặc dù vậy, trong cùng khoảng thời gian đó, khoa học lại cũng có những bước tiến đáng kể trong việc biểu diễn tri thức ngôn ngữ, tiêu biểu là việc hình thành, xây dựng những mạng ngữ nghĩa lớn (semantic networks). Bước tiến đáng kể đó đã thúc đẩy trở lại hoạt động nghiên cứu xử lý ngôn ngữ tự nhiên, mà điển hình là bài toán xử lý nhập nhằng ngữ nghĩa.

Trong khoảng hai thập kỷ tiếp theo, hầu hết các nghiên cứu khử nhập nhằng ngữ nghĩa vẫn chỉ giới hạn trong các mô hình trí tuệ nhân tạo với trí thức do người đưa vào dưới dạng các tập luật. Do đó, những mô hình khử nhập nhằng ngữ nghĩa trong thời điểm này vẫn còn rất hạn chế và chỉ hoạt động tốt trong phạm vi một số mẫu câu nhất định, cụ thể nào đấy.

Đến những năm 90, cùng với sự bùng nổ của công nghệ thông tin, các tài liệu, văn bản điện tử xuất hiện ngày càng đã hỗ trợ rất nhiều cho những nghiên cứu về xử lý nhập nhằng ngữ nghĩa – hướng tiếp cận đến bài toán này đã chuyển từ hướng luật

sang hướng ngữ liệu và đạt được nhiều bước tiến khả quan. Cũng trong khoảng thời gian này, các vấn đề khác của ngôn ngữ tự nhiên có liên quan trực tiếp đến xử lý nhập nhằng ngữ nghĩa, như tách câu, đối sánh văn bản song ngữ, gán nhãn ngữ pháp, phân tích cấu trúc ngữ pháp cũng đã được nghiên cứu và giải quyết khá hoàn chỉnh. Dựa trên nền tảng đó, từ những năm 90 cho tới gần đây, xử lý nhập nhằng ngữ nghĩa hiện là bài toán trọng tâm thu hút rất nhiều các công trình nghiên cứu.

1.3 NHỮNG KHÓ KHĂN, THỬ THÁCH

Những khó khăn, thử thách trong việc xử lý nhập nhằng ngữ nghĩa bao gồm :

- Chúng ta chưa có một định nghĩa rõ ràng về ngữ nghĩa của một từ. Cụ thể, cách duy nhất để chúng ta định nghĩa các sắc thái ý nghĩa của một từ là dựa vào từ điển, thế nhưng bản thân các từ điển cũng chưa nhất quán với nhau do phạm vi của ngôn ngữ quá rộng lớn.

- Giữa các sắc thái ý nghĩa của cùng một từ đôi khi không có một ranh giới rõ ràng để phân biệt, ví dụ như từ *title* có các sắc thái ý nghĩa sau :

    - Tên, tựa đề của một quyển sách, bức tranh, bộ phim hay tác phẩm nghệ thuật …
    - Quyền sở hữu đất đai. (1)
    - Văn bản chứng nhận quyền sở hữu. (2)
    - Danh xưng kèm theo tên.

    Hai sắc thái ý nghĩa (gạch dưới) thật sự rất khó phân biệt ngay cả khi được đặt trong một ngữ cảnh cụ thể và rõ ràng. Những sắc thái ý nghĩa như vậy chỉ có thể được phân biệt dựa vào việc phân tích ngữ dụng (pragmatic use) nhưng bản thân ngữ dụng thì lại thường thay đổi và không nhất quán.

- Ngữ liệu chuyên dùng cho mục đích xử lý nhập nhằng ngữ nghĩa còn quá ít. Vấn đề khử nhập nhằng ngữ nghĩa nói cho cùng chính là việc đi sâu vào bản chất của ngôn ngữ (nature of language). Để có thể nắm bắt được dù chỉ phần nào bản chất của ngôn ngữ, chúng ta vẫn cần phải có một lượng rất lớn ngữ liệu huấn luyện. Tất nhiên, việc thu thập một lượng lớn tài nguyên thô hiện nay không khó nhưng để gán nhãn ngữ nghĩa hoàn chỉnh cho lượng tài nguyên khổng lồ đó thì lại đòi hỏi rất nhiều công sức.

Những khó khăn nói trên đã đem đến rất nhiều trở ngại cho bài toán khử nhập nhằng ngữ nghĩa. Cho đến nay, có thể nói là vẫn chưa có một cách tiếp cận nào cho ra một kết quả hoàn chỉnh. Hầu hết các công trình liên quan vẫn chỉ dừng lại ở mức thí nghiệm do thiếu dữ liệu huấn luyện. Tuy nhiên, những công trình nghiên cứu đó cũng đóng góp rất nhiều ý tưởng rất quan trọng và phần nào giải quyết được bài toán khử nhập nhằng ngữ nghĩa (dù chưa hoàn chỉnh). Mặt khác, trước tình hình ngày càng phát triển của công nghệ thông tin, chúng ta hoàn toàn có quyền hy vọng là những khó khăn nói trên sẽ dần dần được giải quyết để xây dựng một mô hình xử lý nhập nhằng ngữ nghĩa hoàn thiện hơn.

1.4 CÁC HƯỚNG TIẾP CẬN

1.4.1 CÁCH TIẾP CẬN THEO LUẬT

Đây là hướng tiếp cận đầu tiên nhằm khử nhập nhằng ngữ nghĩa được xây dựng dựa trên cơ sở của các phương pháp trí tuệ nhân tạo – là các kỹ thuật chuyển giao tri thức của con người cho máy tính, thường được thể hiện dưới dạng tập luật. Vào đầu thập niên 60, những kỹ thuật chuyển giao tri thức cho máy tính được nghiên cứu áp dụng rất nhiều trong các ứng dụng máy tính thông minh. Một trong số các ứng dụng đó là vấn đề chuyển giao tri thức ngôn ngữ cho máy tính. Từ mục đích ban đầu là

nhằm giúp máy tính hiểu được ngôn ngữ của con người, nhằm phát triển khả năng giao tiếp giữa máy tính - con người, bài toán xử lý nhập nhằng ngữ nghĩa dần dần được hình thành sau nhiều công trình nghiên cứu.

Đây là cách tiếp cận truyền thống xuất phát từ cách làm của các hệ luật phát sinh trong hệ chuyên gia trong lĩnh vực trí tuệ nhân tạo (AI = Artificial Intelligence). Thông thường các hệ luật này được xây dựng bằng tay bởi các chuyên gia xử lý ngôn ngữ tự nhiên. Việc xây dựng một hệ luật như thế đòi hỏi công sức rất lớn và thường không bao quát hết mọi trường hợp, mặc dù, trong một số miền hẹp thì chúng tỏ ra hiệu quả.

Vấn đề thực sự nảy sinh khi chúng ta cần mở rộng quy mô để bao quát hết các hiện tượng của ngôn ngữ. Ban đầu, các nhà chuyên môn cho rằng để mở rộng quy mô của hệ khử nhập nhằng ngữ nghĩa thì ta cứ việc thêm nhiều luật vào; nhưng, thực tế đã cho thấy khi số luật tăng lên thì bản thân người thiết kế sẽ khó mà kiểm soát được tính hợp lý và tương thích của các bộ luật do mình đưa vào vì thế, sẽ xuất hiện nhiều luật mâu thuẫn nhau. Kết quả là những hệ thống khử nhập nhằng ngữ nghĩa được xây dựng trên luật sẽ có nguy cơ bị sụp đổ bởi chính sức nặng của chúng.

1.4.2 CÁCH TIẾP CẬN HƯỚNG NGỮ LIỆU

Do những hạn chế của cách tiếp cận dựa trên luật nói trên, nên trong những năm gần đây, các nhà ngôn ngữ học – máy tính trên thế giới đã chuyển sang cách tiếp cận hướng ngữ liệu. Sự chuyển hướng này cũng xuất phát từ việc ra đời các kho ngữ liệu lớn trên thế giới cùng với sự gia tăng sức mạnh (bộ nhớ, tốc độ, kỹ thuật) của máy tính trong thập niên gần đây. Điểm đặc biệt của cách tiếp cận này là dựa trên cơ sở lý thuyết ngôn ngữ học để học các quy luật của ngôn ngữ tự nhiên từ ngữ liệu.

Trong cách tiếp cận này, máy tính cần có ngữ liệu rất lớn dạng văn bản đơn ngữ, song ngữ hay dạng từ điển (LLOCE, LDOCE, WordNet). Đặc điểm của cách tiếp cận này là nó tự rút ra các quy luật của ngôn ngữ. Nó có những ưu điểm của cách tiếp cận dựa trên luật và đồng thời tránh được những khuyết điểm của việc xây dựng luật thủ công bởi các chuyên gia. Các luật rút ra lại được thử nghiệm tại chỗ để đánh giá độ chính xác (dựa trên ngữ liệu huấn luyện), chính vì thế các luật rút ra tương đối chính xác, bao quát và không mâu thuẫn.

Các phương pháp khử nhập nhằng ngữ nghĩa hướng ngữ liệu thường được phân loại dựa trên cách thức tiếp cận bản chất ngôn ngữ của chúng là có giám sát hay không có giám sát (supervised or unsupervised learning). Với cách tiếp cận có giám sát, chúng ta cần một kho ngữ liệu được gán nhãn ngữ nghĩa hoàn chỉnh (thường là phải chuẩn bị bằng tay), từ đó tiến hành học mẫu để nhận biết và phân loại. Với cách tiếp cận không giám sát, quy trình học có thể được hình dung như một quá trình gom nhóm các mẫu học từ ngữ liệu thô chưa được gán nhãn để từ đó rút ra tri thức.

Một cách phân loại khác dựa trên bản chất tài nguyên sử dụng trong quá trình học máy là hướng ngữ liệu hay hướng từ điển (corpus-based or dictionary-based). Với các cách tiếp cận hướng từ điển, tài nguyên thường được sử dụng là các thể học (ontology) như MRD (machine readable dictionary – LLOCE, LDOCE), WordNet. Nguyên tắc của các cách tiếp cận này nói chung là dựa trên mối liên hệ ngữ nghĩa (synonym, hypernym, hyponym, …) giữa các từ để xây dựng bộ luật hướng ngữ cảnh.

Nói chung thì các phương pháp kể trên đều có ưu điểm và khuyết điểm. Đối với hướng học có giám sát thì ưu điểm là có thể tận dụng rất nhiều mô hình học (có giám sát) tổng quát được phát triển và ứng dụng với độ chính xác cao (khảo sát qua thực nghiệm) và khuyết điểm là nó đòi hỏi một lượng lớn ngữ liệu không những phải gán nhãn hoàn chỉnh mà còn phải được chọn lọc tinh tế cho mục đích sử dụng

(ngữ liệu phải phân bố đều để tránh tình trạng dữ liệu thưa (spareness) ảnh hưởng đến chất lượng học mẫu).

Đối với hướng học không giám sát thì lợi thế là không phải mất nhiều công sức để tinh chế dữ liệu nhưng bù lại thì các mô hình học không giám sát thường bị ảnh hưởng nhiều bởi nhiễu (do thuần túy dựa trên lý thuyết xác suất) và cho kết quả thấp hơn so với các mô hình có giám sát.

Đối với phương pháp học hướng từ điển thì lợi điểm là tài nguyên sử dụng tinh chế, cô đọng dễ sử dụng, giàu thông tin nhưng cách tiếp cận này bất lợi ở chỗ các thể học dùng trong hướng tiếp cận này bị hạn chế (vì mục đích của các loại tài nguyên này chỉ nhằm cung cấp một nguồn tri thức tham khảo) nên không đủ để bao quát bản chất ngôn ngữ. Cụ thể là chúng ta chỉ có thể tiếp cận thông tin ở mức từ vựng và mối liên hệ giữa chúng mà bỏ qua những thông tin ở mức cao hơn (cụm từ, ngữ, câu, …) nên kết quả thực thi thường không được như mong đợi.

Đứng trước thực tế đó, xu hướng hiện nay là kết hợp các phương pháp nói trên để đạt được hiệu quả cao hơn. Thông thường, các hệ thống khử nhập nhằng ngữ nghĩa tự động thường học từ dữ liệu thô (không có nhãn ngữ nghĩa) do ngữ liệu tinh chế không nhiều mà lại khá đắt. Trong các hệ thống như vậy, để nâng cao độ chính xác, người ta thường sử dụng thông tin bổ sung từ các thể học như MRD (LLOCE, LDOCE), WordNet, ngữ liệu song ngữ (đã được đối sánh ở mức câu) để hạn chế bớt nhiễu trong quá trình học không giám sát.

Một cách tiếp cận khác ít phổ biến hơn là áp dụng học có giám sát trên một mẩu nhỏ của dữ liệu tinh chế. Hệ thống sau đó sẽ được áp dụng lên một lượng lớn ngữ liệu thô nhằm rút trích thêm thông tin để tự củng cố, nâng cấp chất lượng thực thi (các thông tin được chọn thường thỏa mãn một ngưỡng tin cậy nhất định nào đấy);

quá trình trên cứ tiếp tục cho đến khi ngưỡng dao động của các tham số hệ thống đủ nhỏ (hội tụ).

Nói tóm lại, trong các hướng tiếp cận gần đây, mô hình phổ biến là kết hợp nhiều loại tài nguyên khác nhau để nâng cao hiệu quả. Nguyên tắc là thay vì tìm kiếm những mô hình phù hợp với một loại tài nguyên nào đấy thì người ta tìm cách thiết kế, cải tiến các mô hình nhằm tích hợp nhiều loại tài nguyên khác nhau để có được chất lượng tốt nhất. Trong các phần sau, chúng ta sẽ lần lượt điểm qua các phương pháp kinh điển bên cạnh những hướng nghiên cứu gần đây để có được cái nhìn bao quát trong lĩnh vực này.

CHƯƠNG 2 – CƠ SỞ LÝ THUYẾT

2.1 CƠ SỞ LÝ THUYẾT NGÔN NGỮ HỌC

2.1.1 KHÁI NIỆM VỀ NHÃN NGỮ NGHĨA CỦA TỪ

Thông qua việc khảo sát ý nghĩa từ vựng của mỗi từ thực, ta thấy về cơ bản thì mỗi từ có thể mang nhiều sắc thái ý nghĩa khác nhau tùy thuộc vào ngữ cảnh sử dụng của chúng. Chẳng hạn, danh từ ”*bank*” trong tiếng Anh có thể là “*ngân hàng*”, hoặc “*bờ sông*” hay “*dãy*”; danh từ “*plant*” trong tiếng Anh có thể là “*thực vật*” hay “*nhà máy*”. Để dễ phân biệt các ngữ nghĩa từ vựng khác nhau, các nhà ngữ nghĩa học, từ vựng học và tâm lý học – ngôn ngữ đã phân chia toàn bộ các ý nghĩa từ vựng có thể có thành hệ thống các ý niệm (cây phả niệm – ontology) và mỗi ý niệm như vậy được coi như là một nhãn ngữ nghĩa của từ.

Để dễ hiểu chúng ta có thể lấy ví dụ với danh từ “*bank*” ở trên. Các sắc thái ý nghĩa tương ứng của nó sẽ là “*ngân hàng*” thuộc về ý niệm “*công trình xây dựng nhân tạo*”; “*bờ sông*” sẽ thuộc về ý niệm “*công trình thiên tạo*”; “*dãy*” sẽ thuộc về ý niệm “*sự sắp xếp tổ chức*”. Tương tự, với danh từ “*plant*”, sắc thái ý nghĩa “*thực vật*” sẽ thuộc nhóm ý niệm “*sự sống*” còn sắc thái ý nghĩa “*nhà máy*” sẽ thuộc nhóm ý niệm “*máy móc, thiết bị*”.

Về mặt tổ chức, từ điển nhãn ngữ nghĩa có tổ chức khác hẳn so với cách tổ chức quen thuộc của các từ điển thông thường, vốn chỉ chú trọng đến tính hợp lý và chặt chẽ về mặt hình thức (hình thái) nhưng lại bỏ qua tính hợp lý về mặt nội dung (ngữ nghĩa) và cũng không phù hợp với tư duy ngôn ngữ của con người. Ví dụ, với tổ chức của từ điển thông thường (đơn ngữ hay song ngữ), các từ được sắp xếp theo thứ tự ABC của mục từ, chính vì vậy mà hai mục từ “*animals*” (động vật) và “*zoo*”

(sở thú), hay "*aunt*" (cô / dì) và "*uncle*" (chú / bác) được đặt ở vị trí rất xa nhau, không phản ánh được mối liên hệ tương đồng về mặt ngữ nghĩa. Qua thực nghiệm, các nhà ngôn ngữ - tâm lý học đã chỉ ra rằng khi đưa ra một từ kích thích, ví dụ như "*aunt*", thì đa số đều cho biết trong đầu họ nghĩ đến từ "*uncle*" trước nhất, điều này chứng tỏ rằng : ngay lời nói bên trong của chúng ta, thì hai từ đó đã có quan hệ gắn kết với nhau. Đây cũng chính là nền tảng lý luận về ngữ nghĩa từ vựng mà các nhà làm từ điển phân lớp ý niệm đã dựa vào khi xây dựng các hệ thống phân lớp ngữ nghĩa và gán nhãn ngữ nghĩa cho các lớp đó.

Hệ thống các ý niệm (concept) này sẽ là chung nhất cho mọi ngôn ngữ, vì hệ thống các ý niệm này được xây dựng dựa trên sự phân chia của thế giới khách quan. Trong khi đó, ngôn ngữ là công cụ của tư duy, mà tư duy là sự phản ánh hình ảnh của thế giới khách quan. Chẳng hạn : khái niệm "*người chồng*" trong tất cả các ngôn ngữ chắc chắn sẽ phải được xây dựng dựa trên các cơ sở ý nghĩa "*người nam*", "*người đã trưởng thành*", "*có gia đình*", "*có vai trò là chồng trong quan hệ với vợ*". Nghĩa là cái biểu đạt trong các ngôn ngữ là khác nhau nhưng cái được biểu đạt thì như nhau vì ý niệm và từ không trùng nhau nên hệ thống ý niệm này có thể được sử dụng cho mọi ngôn ngữ.

Kết quả nghiên cứu về phổ quát ngôn ngữ cũng cho thấy : một số phổ quát ngôn ngữ là từ các hiện tượng tâm lý – ngôn ngữ học, phụ thuộc vào mối quan hệ giữa ngôn ngữ và tư duy của con người. Một số phổ quát ngôn ngữ khác lại là những hiện tượng về dân tộc – ngôn ngữ học, phụ thuộc vào mối quan hệ giữa ngôn ngữ và văn hóa. Các nhà nghiên cứu chia phổ quát ngôn ngữ thành hai dạng sau đây :

- Các phổ quát về thực thể : là những nét chung về sự tổ chức các thực thể ngôn ngữ. Chẳng hạn, mọi ngôn ngữ đều tồn tại các phạm trù danh từ và động từ, nó là cơ sở để biểu hiện cấu trúc chìm của câu trong mọi ngôn ngữ.

- Các phổ quát về dạng thức : chẳng hạn, ngữ pháp tạo simh coi rằng bộ phận cơ sở của cú pháp trong mọi ngôn ngữ thì giống nhau.

Ngoài các phổ quát ngôn ngữ về ngữ âm, ngữ pháp, ngữ nghĩa là những phổ quát chỉ đề cập tới một phương diện ký hiệu hoặc tới cái biểu đạt hoặc tới cái được biểu đạt, người ta còn chú ý tới các phổ quát ngôn ngữ về ký hiệu, chúng đề cập tới cái quan hệ giữa cái biểu đạt và cái được biểu đạt. Trong "Giáo trình ngôn ngữ học đại cương" của Ferdinand de Saussure đã chỉ ra hai dạng quan hệ : ngang (tuyến tính, hình tuyến, ngữ đoạn) và dọc (hệ hình). Tương ứng với quan hệ ngang có trường nghĩa tuyến tính và trường nghĩa biểu niệm. Trường nghĩa biểu vật là tập hợp tất cả những từ đồng nghĩa và ý nghĩa biểu vật, còn trường nghĩa biểu niệm là tập hợp tất cả các từ có chung cấu trúc biểu niệm.

2.1.2 MỘT SỐ HỆ THỐNG NHÃN NGỮ NGHĨA

Nếu như các hệ thống nhãn từ pháp đã được thống nhất và xác định rõ ràng trong các ngôn ngữ (ví dụ, hệ thống nhãn từ pháp Penn Tree Bank (xem Phụ lục A) của tiếng Anh được dùng phổ biến nhất hiện nay), thì ngược lại, việc xây dựng hệ thống nhãn ngữ nghĩa cho đến này vẫn chưa hoàn tất và vẫn đang tồn tại rất nhiều hệ thống nhãn khác nhau. Điểm khó khăn nhất là có những từ ta không biết phân vào ý niệm nào vì cách phân loại còn phụ thuộc vào mục đích và lĩnh vực sử dụng. Ngoài ra, nếu hệ thống nhãn ngữ nghĩa này phân quá chi tiết thì số nhãn sẽ rất lớn (hàng trăm ngàn nhãn) và không thể gán nhãn tự động được (vì khi đó ta sẽ cần đến ngữ liệu huấn luyện có tới hàng tỷ từ). Còn nếu hệ thống nhãn phân quá thô thì nó lại không đáp ứng được một số nhu cầu phần biệt ngữ nghĩa trong thực tế (chẳng hạn, nhu cầu khử mơ hồ trong những trường hợp cùng nhãn ngữ nghĩa nhưng có ý nghĩa từ vựng khác nhau). Trong các phần sau, ta sẽ khảo sát một số hệ thống nhãn ngữ nghĩa thông dụng hiện nay.

2.1.2.1 HỆ THỐNG NHÃN NGỮ NGHĨA LDOCE

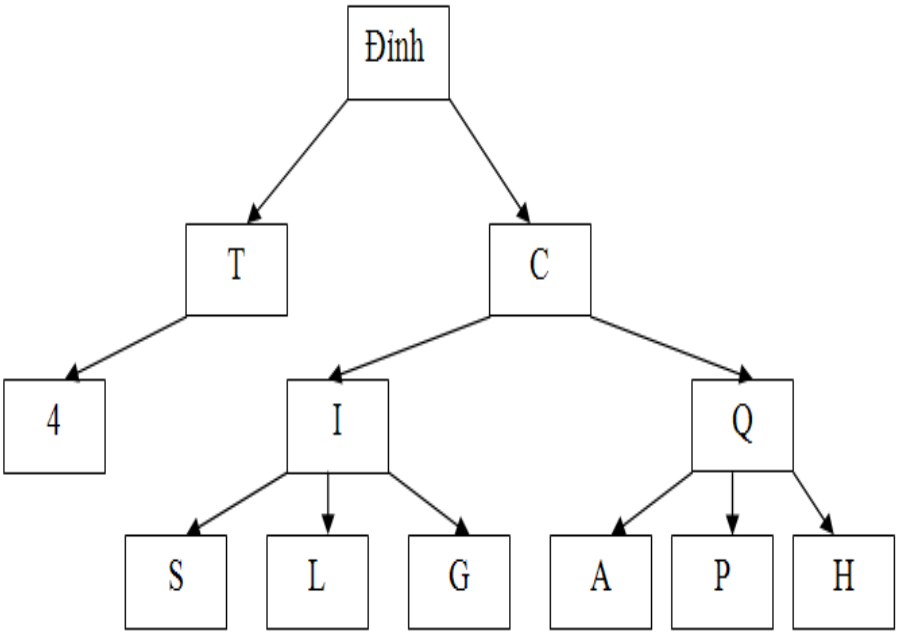
LDOCE (Longman Dictionary Of Contemporary English) gồm 45,000 mục từ với hơn 65,000 nghĩa. Mỗi mục từ được phân biệt dựa trên mã từ loại, mã cú pháp, mã ngữ nghĩa, mã chủ đề, mã phong cách. LDOCE gồm 100 chủ đề chính, như : MD – y học, VH – xe cộ, ON – nghề nghiệp, … Các chủ đề chính có thể được kết hợp với nhau tạo ra các chủ đề con, như MDON – nghề nghiệp / y học. LDOCE gồm 19 mã ngữ nghĩa cơ bản và 13 mã ngữ nghĩa phát sinh được kết hợp từ 19 mã ngữ nghĩa cơ bản trên. Cụ thể, chúng ta có bảng mô tả sau :

STT	Mã ngữ nghĩa cơ bản	Mã ngữ nghĩa phát sinh
1.	A – Con vật (animal)	E – Chất rắn / lỏng (S + L)
2.	B – Con vật cái (female animal)	K – Người / Con vật đực (D + M)
3.	C – Vật cụ thể (concrete)	O – Người / Con vật (A + H)
4.	D – Con vật đực (male animal)	R – Người / Con vật cái (B + F)
5.	F – Người nữ (female human)	U – Tập hợp người / Con vật (Col. + O)
6.	G – Khí (gas)	V – Thực vật / Con vật (P + A)
7.	H – Người (human)	W – Vật cụ thể / trừu tượng (T + I)
8.	I – Vật cụ thể không có sự sống	X – Vật trừu tượng / Người (T + H)
9.	J – Vật rắn di chuyển được	Y – Vật trừu tượng / có sự sống (T + Q)
10.	L – Chất lỏng (liquid)	1 – Người / Chất rắn (H + S)
11.	M – Người nam (male human)	2 – Trừu tượng / Chất rắn (T + S)
12.	N – Vật rắn không di chuyển được	6 – Chất lỏng / Trừu tượng (L + T)
13.	P – Thực vật (plant)	7 – Chất khí / Chất lỏng (G + L)
14.	Q – Có sự sống (animate)	
15.	S – Chất rắn (solid)	
16.	T – Trừu tượng (abstract)	

17.	Z – Không đánh dấu (unmarked)	
18	4 – Vật thể trừu tượng (abs physic)	
19.	5 – Chất hữu cơ (organic material)	

Bảng 1 – Bảng mã ngữ nghĩa của LDOCE

Các mã ngữ nghĩa nói trên có thể được sắp xếp theo cây phân cấp như sau :



Hình 1 – Cây phân cấp các mã ngữ nghĩa của LDOCE

Hầu hết các ngữ nghĩa của danh từ đều mang mã ngữ nghĩa và các mã ngữ nghĩa này được dùng để phân lớp ngữ nghĩa cho danh từ. Đối với động từ và tính từ, các mã ngữ nghĩa này sẽ được dùng để làm tiêu chí chọn ngữ nghĩa cho các đối số (các vai) của các động từ hay tính từ đó. Ví dụ : động từ "*ăn*" cần có danh từ ở vai chủ thể là người / hay con vật (mã O), hay danh từ mà được tính từ "*màu xanh*" bổ nghĩa phải là danh từ thuộc vật cụ thể (mã C), …

	Mục từ	Số nghĩa	Mức đa nghĩa
Danh từ	23,800	37,500	1.6
Động từ	7,921	15,831	1.9
Tính từ	6,992	11,371	1.6
Tổng cộng	38,643	64,702	1.7

Bảng 2 – Thống kê số lượng mức từ, nghĩa của các từ loại trong LDOCE

2.1.2.2 HỆ THỐNG NHÃN NGỮ NGHĨA LLOCE

LLOCE (Longman Lexicon Of Contemporary English) (xem Phụ lục B) là một từ điển ý niệm được xây dựng dựa trên từ điển ý niệm LDOCE. Từ điển LLOCE không sắp xếp các mục từ tiếng Anh theo mẫu tự ABC thông thường, mà sắp xếp thành các chủ đề, mỗi chủ đề được chia thành nhiều nhóm, mỗi nhóm được chia thành nhiều lớp (tạm gọi là lớp ngữ nghĩa) và mỗi lớp gồm các mục từ có quan hệ về nghĩa (nghĩa biểu vật hay nghĩa biểu niệm) với nhau (như : đồng nghĩa, gần nghĩa, …). Tên của mỗi lớp chính là nhãn ngữ nghĩa và các lớp này có mối liên hệ ngữ nghĩa (qua đường kết nối bên trong) với các lớp khác (có thể thuộc chủ đề khác) trong từ điển. Tổng số LLOCE gồm 14 chủ đề, 129 nhóm, 2449 lớp ngữ nghĩa với hơn 16,000 mục từ.

Ví dụ : Chủ đề A là về "*sự sống và vật thể sự sống*" (Life and living things); chủ đề B là về "*cơ thể : chức năng và sự chăm sóc*" (The Body : its Functions and Welfare); chủ đề L là "*không gian và thời gian*"; Chủ đề A được tiếp tục phân thành 10 nhóm sau :

- Sự sống và cái chết, có chứa các lớp từ A1 đến A20.

- Các sinh vật nói chung, có chứa các lớp từ A30 đến A43.

- Động vật và động vật có vú, có chứa các lớp từ A50 đến A61.

- Chim, có các lớp từ A70 đến A78.

- Bò sát và lưỡng cư, có các lớp từ A90 đến A94.

- Cá và các thủy sinh vật khác, có các lớp từ A100 đến A104.

- Côn trùng và các sinh vật tương tự, có các lớp từ A110 đến A113,

- Các bộ phận của động vật, có các lớp từ A120 đến A128.

- Các loài và bộ phận của động vật, có các lớp từ A130 đến A141.

- Thực vật nói chung, có các lớp từ A150 đến A158.

Mỗi lớp ngữ nghĩa trong LLOCE thường gắn với một từ loại cụ thể nào đó và mang một ý nghĩa cụ thể nào đó. Trong mỗi lớp này sẽ chứa một số từ thỏa điều kiện từ loại và ngữ nghĩa chung của lớp. Trong LLOCE sử dụng 3 từ loại chính là danh từ, động từ và tính từ.

Ví dụ :

- Lớp A1 gắn với động từ, có ý nghĩa : "*tồn tại và tạo sự tồn tại*", lớp này bao gồm các động từ sau : exist, be (tồn tại), create (tạo ra), animate (tạo sự sống), ...

- Lớp A2 (động từ), có ý nghĩa : "*sống và chết*", gồm các từ sau : live (sống), live on (tiếp tục sống), exist (tồn tại), die (chết), decay (thối rữa), decompose (phân rã), rot (thối), survive (sống sót), ...

- Lớp A3 (tính từ), có nghĩa là : "*thuộc về sống và chết*", gồm các từ sau : living (đang sống), alive (còn sống), live (sống), animate (có sức sống), dead (chết), dying (sắp chết), ...

- Lớp A4 (danh từ), có nghĩa là : "*sự sống và cái chết*", gồm các từ sau life (đời sống), existence (sự tồn tại), creation (sự tạo ra), animation (sống động), ...

- Lớp A150 (N) : táo, mơ, đào, thơm, dứa, lê, mận, đu đủ, anh đào, nho, xoài, chà là, vả, lựu, ... (trái cây)

- Lớp G148 (N) : chữ cái, mẫu tự, ký tự, chữ hoa, chữ thường, ... (chữ cái)

- Lớp G155 (N) : thư, thư dài và quan trọng, thư ngắn, bản ghi chép, phong bì, bao thư, nhãn, ... (thư từ, ghi chú), ...

Mỗi lớp thường được liên kết chéo (cross – reference) với các lớp ngữ nghĩa khác theo các quan hệ logic – ngữ nghĩa. Ngoài các nhãn về ngữ nghĩa nói trên, LLOCE còn chứa đựng các nhãn về từ loại và cú pháp. Chính các nhãn ngữ pháp này sẽ giúp chúng ta rất nhiều trong việc khử mơ hồ ngữ nghĩa của từ vì nghĩa của từ cũng phụ thuộc rất nhiều vào vai trò ngữ pháp của nó trong câu.

2.1.2.3 HỆ THỐNG NHÃN NGỮ NGHĨA WORDNET

WordNet là một hệ cơ sở tri thức khổng lồ về ngữ nghĩa của từ vựng tiếng Anh với hơn 100,000 ý niệm khác nhau, được xây dựng bởi các nhà ngôn ngữ học – máy

tính, ngôn ngữ học – tâm lý và ngôn ngữ học – tri nhận ở Đại học Princeton (Mỹ) từ đầu thập niên 1980. Hệ WordNet là một hệ trực tuyến (online) cho phép mọi người ở khắp nơi được tự do (miễn phí) khai thác hay tải xuống (download) máy cá nhân của mình cho các mục đích nghiên cứu.

WordNet là một kho tàng trí thức ngữ nghĩa từ vựng khổng lồ và đã được rất nhiều các nhà ngôn ngữ học và ngôn ngữ học – máy tính khai thác, ứng dụng thành công trong nhiều bài toán về xử lý ngữ nghĩa. Hiện nay, WordNet đang được các nhà khoa học về ngôn ngữ, tâm lý, máy tính trên toàn thế giới tiếp tục khai thác đóng góp để cải tiến ngày càng hoàn thiện hơn. WordNet có nhiều ưu điểm không thể chối cãi, đó là : tính khoa học, tính hệ thống, tính mở, tính dễ sử dụng, tính phổ thông, tính phát triển ... Chính vì vậy, đến nay, đã có một số công trình bản địa hóa (localization) WordNet theo ngôn ngữ của một số nước, như : Pháp, Nhật, Tây Ban Nha, Hàn, Hoa ... và gần đây là Việt Nam.

WordNet không chỉ đơn thuần là nhóm các từ đồng nghĩa hay các từ có quan hệ ngữ nghĩa với nhau thành từng lớp như một số từ điển LDOCE, LLOCE, ... mà WordNet còn là một hệ thống các ý niệm có quan hệ nhiều mặt với nhau, tạo thành một mạng lưới phức tạp. Mục tiêu cơ bản của WordNet là chứa các thông tin về ngữ nghĩa của từ, mà hễ nói đến khái niệm hay định nghĩa về từ thì chắc chắn lại dẫn đến nhiều ý kiến khác nhau. Chính vì vậy, ngay từ đầu, ta phải xác định cách hiểu về đơn vị từ trong WordNet là như thế nào, sau đó ta tìm hiểu về tập đồng nghĩa (synset) – một thành phần cơ bản của WordNet để áp dụng vào việc bản địa hóa WordNet thành ngôn ngữ của chúng ta.

2.1.2.4 HỆ THỐNG NHÃN NGỮ NGHĨA CORELEX

Dù WordNet là một nguồn thông tin ngữ nghĩa từ vựng vô cùng phong phú và có giá trị cho hầu hết các bài toán xử lý ngữ nghĩa trong ngôn ngữ tự nhiên, nhưng

WordNet có thiếu sót lớn nhất chính là nó đã bỏ qua sự phân biệt về nguyên tắc giữa từ đa nghĩa (polysemy) với từ đồng nghĩa (homonymy). Để khắc phục thiếu sót đó, trong công trình nghiên cứu của mình, Paul Buitelaar đã đề ra một hệ thống nhãn ngữ nghĩa mới dựa trên nguyên lý là đối với một từ đa nghĩa thì sẽ có những nghĩa không liên quan đến nhau (contrastive) và những nghĩa liên quan một cách hệ thống đến nhau (complementary). Đó chính là hệ thống nhãn ngữ nghĩa CoreLex.

Ta có thể coi các nghĩa không liên quan đến nhau như là các nghĩa của từ đồng tự (homograph), còn các nghĩa có liên quan hệ thống đến nhau là các nghĩa của từ đa nghĩa (polysemy). Ví dụ : "*bank*" (ngân hàng) và "*bank*" (bờ) là hai từ đồng tự, còn "*line*" (dây) và "*line*" (đường) là nhựng nghĩa của từ đa nghĩa. Khác với WordNet, CoreLex chú trọng đến các nghĩa của từ đa nghĩa vì đây là những nghĩa có liên quan hệ thống đến nhau.

Để xây dựng được hệ thống nhãn ngữ nghĩa của CoreLex, Paul Buitelaar đã phân tích các nét ngữ nghĩa của từng danh từ (tổng số 40,000 danh từ), rồi đưa về các nhãn ngữ nghĩa cơ bản. Các nhãn ngữ nghĩa cơ bản này chính là các lớp synset nguyên thủy của WordNet cộng với phần mở rộng. Một danh từ có thể có một hay nhiều nhãn ngữ nghĩa cơ bản. Những nghĩa có liên quan đến nhau sẽ giống nhau ở một nét nghĩa cơ bản nào đó. Tập hợp các nét nghĩa cơ bản giống nhau của một số danh từ, hình thành hệ thống ngữ nghĩa CoreLex với tổng cộng 126 lớp ngữ nghĩa.

2.1.3 NHẬN XÉT CÁC HỆ THỐNG NHÃN NGỮ NGHĨA

Qua khảo sát các hệ thống nhãn ngữ nghĩa của LLOCE, LDOCE, WordNet và CoreLex, có những điểm đáng chú ý sau :

- Cách phân chia các lớp của LLOCE thực chất là dựa trên cơ sở lý thuyết phân chia trường ngữ nghĩa theo trục dọc (trường nghĩa biểu vật và biểu niệm). Đối

với WordNet, ngoài việc dựa trên cơ sở lý thuyết phân chia theo trường biểu vật và biểu niệm, nó còn dựa theo cơ sở phân chia theo trường nghĩa tuyến tính và trường nghĩa liên tưởng (qua các quan hệ chức năng, bộ phận, tính chất).

- Với mục tiêu ban đầu là hệ thống các ý niệm chung nhất cho mọi ngôn ngữ của nhân loại, nên việc biểu diễn hệ thống các ý niệm trong WordNet được dựa trên cơ sở lý thuyết về ngôn ngữ học – tri nhận (cognitive linguistic), ngôn ngữ học – tâm lý (psycho – linguistics), … nhưng tất cả những lý thuyết này đều hướng tới một mục tiêu chung là nghiên cứu về sự chung nhất của mọi ngôn ngữ trên thế giới hay còn gọi là phổ quát (universal) của ngôn ngữ.

- Hệ thống nhãn LDOCE chỉ chú trọng đến danh từ, có số lượng từ khá lớn (45,000) nhưng sự phân chia theo lớp ngữ nghĩa quá thô (chỉ có 32 lớp) nên không đủ sức khử nhập nhằng ngữ nghĩa cho các từ cùng lớp nhưng khác nghĩa.

- Hệ thống nhãn LLOCE có ưu điểm là đơn giản, hệ thống phân cấp chỉ gồm 3 cấp (chủ đề - nhóm – lớp), số nhãn không quá lớn (chỉ gồm 2449 nhãn). Bên cạnh đó, số lượng từ của LLOCE cũng còn hạn chế (chỉ gồm 16,000 mục từ), nên nếu muốn áp dụng vào một hệ thống thực tế, cần phải mở rộng thêm.

- Hệ thống nhãn của WordNet rất chi tiết, đầy đủ (cho các từ loại chính) vì vậy số lượng nhãn rất lớn (hơn 100,000 nhãn), WordNet có ưu điểm là phân cấp chi tiết (hàng chục cấp) và giữa các lớp synset còn có nhiều kiểu quan hệ khác nhau. Chức năng chính của hệ thống nhãn ngữ nghĩa trong đa số các hệ xử lý ngôn ngữ tự nhiên là để khử nhập nhằng ngữ nghĩa ở mức cần thiết chứ không phải cho mục đích hiểu (cần có tri thức chi tiết về thế giới thực) nên không cần phải phân giải ngữ nghĩa chi tiết như trong WordNet. Với số lượng nhãn quá lớn như trong WordNet, thì chúng ta không thể xây dựng được đủ ngữ liệu huấn luyện tổng quát cho tất cả các từ (cần ngữ liệu hàng tỷ từ).

- Hệ thống nhãn CoreLex được xây dựng từ các lớp cơ bản của WordNet, có khả năng phân biệt được từ đồng nghĩa (homonym) và từ đồng tự (homograph) trong khi đó WordNet thì không. Ngoài ra, CoreLex chỉ bao gồm 39 nhãn cơ bản và 126 lớp dẫn xuất với khoảng 40,000 danh từ nên khó áp dụng vào một hệ thống mới với các danh từ không có trong danh sách đó.

Kết luận : Tùy vào mục đích sử dụng mà ta có thể chọn các hệ thống nhãn ngữ nghĩa cho phù hợp :

- Nếu để phân tích và hiểu sâu, ta nên sử dụng hệ thống nhãn WordNet.

- Để khử nhập nhằng ở mức tương đối ta có thể dùng LLOCE hay LDOCE.

- Để gán nhãn ở mức độ thô, dễ hiểu, dễ nhơ, nên dùng nhãn CoreLex và các nhãn ở tầng sơ cấp (primitives) của WordNet.

2.1.4 CÁC NGUỒN TRI THỨC ĐỂ XỬ LÝ NGỮ NGHĨA

Để xử lý ngữ nghĩa, người ta phải kết hợp nhiều nguồn tri thức : tri thức ngôn ngữ (hình thái, ngữ pháp, ngữ nghĩa) và tri thức ngoài ngôn ngữ (tri thức về thế giới thực). Các nguồn tri thức đó thường bao gồm :

2.1.4.1 TRI THỨC VỀ TỪ LOẠI

Trong trường hợp các từ đồng tự (homograph) và có nghĩa khác nhau với các từ loại khác nhau và ứng với một từ loại chỉ có một nghĩa duy nhất, thì nhờ thông tin từ loại, chúng ta sẽ xác định được chính xác nghĩa của chúng. Ví dụ : từ "*can*" có

nghĩa là "*có thể*" (trợ động từ), "*cái hộp*" (danh từ), "*đóng hộp*" (động từ). Vì vậy, với các trường hợp này, nếu biết được chính xác từ loại, chúng ta hoàn toàn khử được nhập nhằng ngữ nghĩa của chúng.

Theo thống kê trong từ điển LLOCE, có tới 88% mục từ thuộc dạng nói trên và 7% mục từ (tập các từ đồng tự) có nhiều từ loại, mỗi từ loại có thể có nhiều nghĩa khác nhau, nhưng trong đó có ít nhất một từ loại có duy nhất một nghĩa. Đối với trường hợp này, ta có thể khử nhập nhằng ngữ nghĩa nếu từ loại của nó (trong ngữ cảnh) chính là từ loại mà chỉ có một nghĩa.

2.1.4.2 TRI THỨC VỀ QUAN HỆ CÚ PHÁP VÀ RÀNG BUỘC NGỮ NGHĨA

Đối với các trường hợp cùng từ loại nhưng có nhiều hơn một nghĩa thì thông tin từ loại không đủ để xử lý nhập nhằng ngữ nghĩa. Ví dụ : từ "*bank*" có hai từ loại là danh từ và động từ. Với danh từ, ta có các ngữ nghĩa "*ngân hàng*", "*bờ sông*", … Trong trường hợp này ta cần phải sử dụng thêm các tri thức về thế giới thực thông qua các ràng buộc ngữ nghĩa (selectional restriction) giữa các thành phần cú pháp (S – V – O – M) trong câu.

2.1.4.3 TRI THỨC VỀ NGÔN TỪ

Sự ràng buộc về ngữ nghĩa giữa các thành phần cú pháp không phải lúc nào cũng giải quyết được mọi nhập nhằng, vì có những quan hệ tiềm ẩn về logic, về ngữ nghĩa hay thậm chí do thói quen mà việc nhận biết phải đòi hỏi những tri thức thế giới thực mà đến nay người ta cũng không thể tích hợp hết vào từ điển hay các cơ sở trí thức khác trong máy tính.

Ví dụ : Danh từ "*bank*" trong câu "*I go to the bank ...*" có nghĩa gì : "*ngân hàng / bờ (sông) / dãy*" ? Rõ ràng nếu chỉ xét đến các yếu tố ngữ pháp thì ta không có cách gì nhận biết được ngữ nghĩa của từ "*bank*" trong câu này.

Vì vậy, để khử nhập nhằng trong các trường hợp này, người ta thường xét đến hình thái và ngữ nghĩa của các từ lân cận hay còn gọi là ngôn từ (collocation). Chẳng hạn, khi thấy "*bank*" đi cùng với "*river*" thì ta biết ngay là đang nói về "*bờ sông*" còn nếu "*bank*" đi cùng với "*account, money*" thì đấy là đang nói về "*ngân hàng*". Thông tin về các từ có quan hệ ngữ nghĩa như trên có thể tìm thấy trong các từ điển dạng Thesaurus của Roget hoặc LLOCE. Khi đó, phạm vi lân cận của từ cần khử ngữ nghĩa có thể là bên trái 1, 2 hay n từ và bên phải 1, 2 hay n từ.

2.1.4.4 TRI THỨC VỀ CHỦ ĐỀ

Trong một số trường hợp nhập nhằng, chúng ta có thể xác định được nghĩa đúng của từ nếu ta biết được chủ đề của văn bản. Lấy ví dụ, từ "*bank*", nếu đang nói về lĩnh vực "*tài chính*" thì nó thường có nghĩa là "*ngân hàng*"; từ "*driver*" có nghĩa là "*trình điều khiển*" nếu chủ đề là lĩnh vực tin học; … Để xác định được chủ đề của văn bản đang cần dịch, ta cần xem xét sự xuất hiện của một số từ chuyên môn trong lãnh vực đó.

Chẳng hạn, nếu trong văn bản ta thấy xuất hiện các từ như "*ellipsis*" (tỉnh lược), "*bilingual*" (song ngữ), "*anaphora*" (thế đại từ), "*phrase*" (ngữ), … thì ta có thể đoán nhận văn bản này đang nói về chủ đề "*ngôn ngữ học*"; tương tự cho các từ "*computer*", "*memory*", "*peripherals*", "*CPU*", … thì chủ đề có thể là "*tin học*"; … Chính vì vậy, trong từ điển LLOCE hay LDOCE đều có mã số chủ đề cho các từ chuyên môn này. Chúng ta có thể xác định được chủ đề một cách tự động bằng cách xem xét các từ chuyên môn lân cận từ đang cần xử lý nhập nhằng ngữ nghĩa gần với chủ đề nào nhất.

2.1.4.5 TRI THỨC VỀ TẦN SUẤT NGHĨA CỦA TỪ

Ta có nhận xét là không phải từ nào cũng thuộc về một chủ đề nào đó (trong từ điển LDOCE, hơn 56% từ thuộc dạng này), vì vậy tính thông dụng của một nghĩa nào đó còn được dựa trên độ đo về tần suất (frequency) xuất hiện của từ đó với nghĩa cụ thể đó. Chẳng hạn, danh từ "*pen*" sẽ có nghĩa thông dụng nhất là "*bút/viết*" (bên cạnh các nghĩa ít thông dụng hơn, như "*chuồng*", "*lông chim*"); "*ball*" có thường có nghĩa là "*quả banh/hòn bi*" hơn là "*buổi khiêu vũ*", …

Độ đo tần suất xuất hiện của mỗi nghĩa của mỗi từ được thống kê trên những ngữ liệu rất lớn thuộc nhiều loại văn bản khác nhau. Chính vì vậy, trong WordNet và trong LDOCE, các nghĩa được sắp xếp theo thứ tự giảm dần (nghĩa thông dụng nhất sẽ được liệt kê đầu tiên).

2.1.4.6 TRI THỨC TRONG ĐỊNH NGHĨA CỦA NGHĨA TỪ (DEFINITION)

Trong các từ điển LDOCE / WordNet, mỗi nghĩa sẽ được định nghĩa và ví dụ kèm theo. Ví dụ : từ "*bank*" trong LDOCE sẽ có các nghĩa kèm định nghĩa của nó như :

- "*land along the side of a river, lake, etc.*" (đất dọc bên sông / hồ)
- "*a place where money is kept and paid …*" (nơi giữ tiền và trả tiền)
- "*a row, a line of …*" (một hàng, một dãy, …)

Dựa trên thông tin trong các định nghĩa này, và so sánh với thông tin của ngữ cảnh, ta có thể xác định được nghĩa phù hợp của từ trong ngữ cảnh đó. Để thực hiện điều

này, Wilks et al. đã tính toán phần giao (overlap) của tất cả các tổ hợp nghĩa của các từ thực trong câu tiếng Anh dùng để định nghĩa mỗi nghĩa của từ.

2.1.5 CÁC MỨC ĐỘ NHẬP NHẰNG TRONG XỬ LÝ NGỮ NGHĨA

2.1.5.1 NHẬP NHẰNG MỨC TỪ VỰNG

Xét câu ví dụ "*I enter the bank*" ở trên, sau khi phân tích cú pháp, máy tính đã xác định được mối quan hệ giữa động từ "*enter*" (đi vào) và đối từ của nó là danh từ "*bank*", nhưng để chọn nghĩa thích hợp cho từ "*bank*" (ngân hàng hay bờ sông) thì phải phân tích nghĩa của động từ "*enter*" và danh từ "*bank*". Trong trường hợp này máy sẽ vận dụng các ý niệm của ngôn ngữ học tri nhận để biết rằng "*enter*" là hành động đi vào không gian kín trong khi bờ sông có ý nghĩa là không gian hở nên không thích hợp với ngữ cảnh. Từ đó, máy tính có thể ra quyết định rằng ngữ nghĩa của từ "*bank*" trong trường hợp này là ngân hàng.

2.1.5.2 NHẬP NHẰNG MỨC CẤU TRÚC

Xét ngữ "*Old man and woman*", về mặt cấu trúc thì ta có hai cách phân tích : *Old [man and woman]* và *[Old man] and [woman]*. Trong trường hợp này thì cách phân tích thứ nhất là phù hợp nhất do tính cân bằng của liên từ *and*. Tuy nhiên, nếu xét ngữ "*Old man and child*" thì ta chỉ có thể phân tích theo cách thứ hai *[Old man] and [child]* vì có sự đối lập giữa thuộc tính trẻ trong từ "*child*" và thuộc tính già trong "*man*".

Về nhập nhằng cấu trúc, chúng ta còn gặp trường hợp quen thuộc sau : "*the man saw the monkey with a telescope*". Trong câu này, máy không biết là giới ngữ *[with a telescope]* sẽ bổ nghĩa cho danh từ "*monkey*" hay cho động từ "*saw*" là "*con*

người" nên giới ngữ "*with a telescope*" có khả năng đóng vai "*dụng cụ cách*" (instrument) cho động từ "*saw*" nhiều hơn. Đó cũng chính là vấn đề nhập nhằng giới ngữ thường gặp.

2.1.5.3 NHẬP NHẰNG MỨC LIÊN CÂU

Xét ví dụ "*The monkey ate the banana because it was hungry*". Trong trường hợp này, máy tính phải xác định được đại từ "*it*" thay thế cho từ nào : "*monkey*" hay "*banana*". Tương tự như từ "*them*" trong câu "*The nurses keep clean sheets and blankets in them*". Máy không biết được đó là người hay vật nếu không nhờ tới ngữ nghĩa của câu phụ trợ trước nó : "*The room has two cabinets. The nurses keep … in them.*" Đó cũng chính là vấn đề nhập nhằng thế đại từ (anaphora).

2.1.5.4 NHẬP NHẰNG MỨC NGỮ DỤNG

Trong một số trường hợp nhập nhằng ngữ nghĩa, ta không thể dùng thông tin trong nội bộ câu hay câu lân cận mà phải xét trên toàn bộ nội dung văn bản xem vấn đề chính là gì, thuộc lĩnh vực nào để từ đó có thể chọn đúng nghĩa của từ / câu. Lấy ví dụ đơn giản ở từ "*sentence*" vừa có nghĩa là câu vừa có nghĩa là án tù. Như vậy, để quyết định đúng sắc thái ý nghĩa cho từ "*sentence*" thì ta phải xem xét xem văn bản này đang bàn về "*ngữ pháp*" hay "*luật pháp*". Để xác định được ngữ cảnh của toàn bộ văn bản, máy tính phải dựa vào sự xuất hiện của các từ khóa (keyword) trên toàn bộ văn bản.

2.2 CƠ SỞ LÝ THUYẾT TIN HỌC

2.2.1 XÂY DỰNG KHO NGỮ LIỆU

Trong phần này, chúng ta sẽ nói sơ về quy cách xây dựng kho ngữ liệu. Kho ngữ liệu ở đây, như đã đề cập ở phần mở đầu, là ngữ liệu song ngữ và đơn ngữ. Theo cách tiếp cận trình bày trong luận văn này, ngữ liệu song ngữ cần phải được đối sánh ở mức câu còn ngữ liệu đơn ngữ cần được phân ra theo từng văn bản (có nội dung nhất quán). Ở đây, bài toán mà chúng ta quan tâm là xử lý nhập nhằng ngữ nghĩa trong dịch máy Anh – Việt nên dĩ nhiên ngữ liệu song ngữ mà chúng ta đề cập ở đây là ngữ liệu song ngữ Anh – Việt, còn ngữ liệu đơn ngữ là tập hợp các văn bản tiếng Anh đã được tách ra thành từng câu.

Như vậy, vấn đề chủ yếu của việc xây dựng kho ngữ liệu cho xử lý nhập nhằng ngữ nghĩa liên quan đến ngữ liệu song ngữ Anh – Việt vì việc chuẩn bị ngữ liệu đơn ngữ tiếng Anh căn bản không có gì đáng nói, do những nguồn tài nguyên như vậy được công bố miễn phí trên nhiều website về xử lý ngôn ngữ tự nhiên.

Đối với ngữ liệu đơn ngữ, chúng ta có thể thu thập dữ liệu từ nhiều nguồn như :

- Nguồn Internet : đây là nguồn ngữ liệu khổng lồ, đã tồn tại sẵn dưới dạng điện tử (nên không phải nhập liệu lại bằng tay). Kho này có vô vàn các lĩnh vực / phong cách khác nhau (cần lọc lại).

- Nguồn sách điện tử (E-books) : bao gồm các sách chuyên ngành khác nhau, như : Tin học, Điện tử, Kinh tế, …

- Nguồn từ điển : trong mỗi từ điển, ở mỗi mục từ, thường chứa các ví dụ mẫu hướng dẫn sử dụng từ đó. Ngôn ngữ trong từ điển là đúng chuẩn ngôn ngữ. Nội dung trong từ điển cũng rất phong phú bao quát.

- Ngữ liệu huấn luyện : đây là những kho ngữ liệu điện tử (thường là tiếng Anh) được xây dựng bởi các nhà ngôn ngữ học – máy tính nước ngoài, như PTB (Penn Tree Bank), SUSANNE, …

Đối với ngữ liệu song ngữ, hiện nay đã có rất nhiều nguồn ngữ liệu điện tử của các tổ chức dịch ngữ liệu điện tử quốc tế được dịch ra nhiều thứ tiếng. Tuy nhiên, điểm bất lợi là các bản dịch đó thường là dịch thoát ý, dịch ý chính, không dịch 1 – 1 (nhất là những văn bản không phải thuộc lĩnh vực khoa học – kỹ thuật). Nói tóm lại, ngữ liệu song ngữ tinh chế không thể thu được đơn giản từ Internet. Trong nghiên cứu này, bộ ngữ liệu song ngữ đem vào sử dụng là ngữ liệu EVC của nhóm VCL. Ngữ liệu song ngữ này bao gồm 400, 000 cặp câu dịch Anh – Việt đã được thu thập, tinh chỉnh qua nhiều bước và thỏa mãn những tiêu chuẩn sau đây :

- Chuẩn ngôn ngữ : Ngữ liệu đều là những văn bản với những câu được xem là chuẩn mực, nghĩa là đúng ngữ pháp và thông dụng. Những văn bản hay bản dịch có tính cá nhân sẽ không được xem xét do không đáp ứng tính thực tế của ngữ liệu.

- Phong cách và lĩnh vực của ngữ liệu : Tiêu chuẩn này tùy thuộc vào mục đích nghiên cứu. Nếu thu thập ngữ liệu để xây dựng từ điển tần số hay phân loại văn bản thì chúng ta cần thu thập nhiều lĩnh vực, phong cách khác nhau. Nếu thu thập để huấn luyện xử lý tự động, ta chỉ cần giới hạn trong một lĩnh vực cụ thể của khoa học kỹ thuật, chứ không nên chọn những lĩnh vực kiểu như văn học (vì lĩnh vực này đến nay máy tính vẫn chưa thể xử lý tự động được).

- Dung lượng và độ phong phú của ngữ liệu : Đơn vị thu thập ngữ liệu phải là văn bản (văn bản không đơn thuần là tập hợp các câu mà là một hệ thống các câu). Độ dài của một văn bản nên ở mức trung bình (khoảng vài ngàn từ, như PTB, SUSANNE chọn khoảng 2000 từ / văn bản). Kho ngữ liệu thu thập được phải

chứa hầu hết (hơn 80%) vốn từ, số lượng kết cấu ngữ pháp khác nhau trong một hay nhiều lĩnh vực nghiên cứu.

- Cách dịch 1 – 1 : Riêng với các ngữ liệu song ngữ, chúng phải thực sự là bản dịch 1 – 1 của nhau, không dịch thoát ý, tóm lược, tương đương hay dịch theo kiểu giải thích diễn giải. Lý do là nếu không phải là dịch 1 – 1, thì máy tính rất khó liên kết từ một cách tự động cho song ngữ đó được. Ngoài ra, bản dịch 1 – 1 còn cần thiết để có thể so sánh, đối chiếu trên từng cấp độ giữa hai ngôn ngữ.

- Ngữ liệu dạng điện tử : Ngoài ba tiêu chuẩn bắt buộc trên, chúng ta sẽ ưu tiên chọn những ngữ liệu nào đang tồn tại dưới dạng điện tử, hoặc có thể chuyển tự động tương đối về dạng điện tử (như các sách in còn rõ), như vậy đỡ tốn công sức nhập liệu bằng tay vào máy tính.

2.2.2 LIÊN KẾT TỪ TRONG NGỮ LIỆU SONG NGỮ

Một từ (hay một đơn vị ngôn ngữ nào đó) thường mang nhiều hơn một nhãn hình thái / ngữ pháp / ngữ nghĩa, … Vì vậy, vấn đề khó khăn nhất trong việc gán nhãn ngôn ngữ chính là làm thế nào để chọn được nhãn đúng trong số các nhãn khả dĩ của một đơn vị ngôn ngữ ? Đây chính là bài toán khử tính mơ hồ vốn có của ngôn ngữ tự nhiên ở hầu hết các cấp độ (từ, ngữ, câu) ở các khía cạnh (hình thái, ngữ pháp, ngữ nghĩa, ngữ dụng). Đây là công việc khó khăn, tốn kém thời gian và công sức nhất. Lấy ví dụ từ kho ngữ liệu Penn Tree Bank của tiếng Anh, người ta đã phải mất hàng chục năm với chi phí hàng triệu USD để gán nhãn ngữ pháp cho 4.5 triệu từ của kho ngữ liệu tiếng Anh PTB.

Vì vậy, để tiết kiệm công sức, các nhà ngôn ngữ học – máy tính đã tìm cách tiếp cận thông qua các mô hình toán học để có thể gán nhãn ngữ liệu một cách tự động. Riêng đối với ngữ liệu song ngữ, việc gán nhãn tự động có thể được thực hiện thông

qua quá trình giải quyết bài toán liên kết từ nhằm liên kết một từ trong ngôn ngữ này với từ / ngữ tương ứng trong ngôn ngữ kia. Khi đã có mối liên kết từ, ta có thể lợi dụng thông tin trong ngôn ngữ này để khử nhập nhằng trong việc gán nhãn ngôn ngữ với độ chính xác cao hơn vì thông thường thì những ngữ nghĩa khác nhau của từ đa nghĩa trong ngôn ngữ nguồn thường có xu hướng thể hiện khác nhau trên ngôn ngữ đích.

Về bản chất, liên kết từ trong ngữ liệu song ngữ là liên kết một từ e_i trong ngôn ngữ *E* với từ v_i trong ngôn ngữ *V*. Tất nhiên, mối liên hệ đó không thể là ánh xạ 1 – 1 mà phải có dạng m – n vì có sự khác biệt về loại hình giữa *E* và *V*. Do đó, bài toán liên kết từ không chỉ đơn thuần là tra từ điển song ngữ như ta nghĩ mà còn phải xét đến những yếu tố về sự từ vựng hóa, phương tiện ngữ pháp / từ vựng, trật tự từ, …

Ví dụ :
Xét cặp câu dịch Anh – Việt sau :
(E) : *Jet planes fly about nine miles high.*
(V) : *Các máy bay phản lực bay cao khoảng chín dặm.*

Nếu dùng từ điển song ngữ thông thường để liên kết câu thì trong từ điển những từ như "*máy bay*", "*mặt phẳng*" sẽ được tìm thấy trong chỉ mục của *plane*. Từ đấy máy tính lại tìm kiếm tiếp trong câu tiếng Việt và chỉ thấy từ "*máy bay*" và do đó máy tính kết luận là từ *planes* được liên kết với từ "*máy bay*". Tuy nhiên, chuyện gì sẽ xảy ra nếu *planes* được dịch là "*phi co*" trong câu tiếng Việt ? Khi đó, rất có khả năng là trong từ điển song ngữ không có chứa từ "*phi co*" và như vậy, máy tính sẽ không thể liên kết từ một cách chính xác được.

Qua sự phân tích ở trên chúng ta thấy là việc liên kết từ không thể chỉ đơn thuần dựa vào từ điển song ngữ, vì trong thực tế dịch, một từ có nhiều cách dịch khác nhau mà trong từ điển không lưu trữ (lý do là vì từ điển chỉ đóng vai trò tài nguyên

tham khảo – do đó, nó chỉ lưu trữ những cách dịch phổ biến để chuyển tải ý niệm rõ ràng nhất cho người dùng). Để giải quyết vấn đề này, hiện chúng ta có hai mô hình liên kết từ phổ biến là mô hình SC (Semantic Class) – Phân lớp ngữ nghĩa (S.J.Ker và J.S.Chang [13]) và mô hình SMT (Statistical Machine Translation) – Liên kết từ thống kê (Brown et al. [7]). Nói một cách khách quan thì ưu điểm của mô hình SC là các mối liên hệ có được thì chính xác (do xét đến nghĩa của từ), nhưng khuyết điểm là có nhiều từ chức năng, hư từ hay các trường hợp thành ngữ dịch thoáng thì mô hình này không liên kết được. Trong khi đó, mô hình SMT của Brown thì có ưu điểm là liên kết được tất cả các từ (về mặt lý thuyết) mà không cần biết nghĩa của từ nhưng lại tạo ra nhiều mối liên hệ không chính xác. Trong các phần sau, chúng ta sẽ lần lượt khảo sát hai cách tiếp cận này.

2.2.2.1 LIÊN KẾT TỪ BẰNG LỚP NGỮ NGHĨA

Mô hình SC đề xuất bởi (S.J. Ker và J.S. Chang [13]) về nguyên tắc bao gồm 3 thuật toán con được thực hiện nối tiếp nhau. Đầu tiên, mô hình liên kết từ bằng từ điển song ngữ sẽ được triển khai để tạo ra danh sách các liên kết ứng viên. Sau đó, mô hình liên kết các lớp ngữ nghĩa dựa trên từ điển đồng nghĩa được triển khai nhằm xây dựng thông tin bổ sung để dựa vào đó, loại bỏ bớt những liên kết ứng viên kém triển vọng. Trong mô hình liên kết từ cuối cùng, yếu tố trật tự từ sẽ được xem xét nhằm đưa ra danh sách liên kết từ tối ưu. Trước khi đi sâu vào chi tiết của từng thuật toán nói trên, ta sẽ thống nhất lại các ký pháp sử dụng để tiện cho việc theo dõi.

Ta quy ước :

- Ngữ liệu song ngữ là *C* bao gồm nhiều cặp câu (*S, T*) tương ứng là câu nguồn và câu đích.

- Gọi *s* là từ hay cụm từ trong *S* và *t* là từ hay cụm từ được dịch theo ngữ cảnh trong *T*.

- Gọi DT_s là tập các nghĩa (được thể hiện trên ngôn ngữ đích) trong từ điển song ngữ của mục từ *s*, mỗi một nghĩa riêng biệt được ký hiệu là *dt*.

- Gọi *CX*, *CY* lần lượt là tập hợp các lớp ngữ nghĩa (tập đồng nghĩa) của ngôn ngữ nguồn và ngôn ngữ đích.

- $W_T = \{ wt \mid wt \in T \wedge wt \in VD \}$ với *VD* là tập các từ trong ngôn ngữ đích.

- $W_S = \{ s \mid s \in S \wedge s \in SD \}$ với *SD* là tập các từ trong ngôn ngữ nguồn.

2.2.2.1.1 THUẬT TOÁN DICT-ALIGN

- Bước 1 : Phân tích *S* (tách từ nếu cần thiết) để thu được danh sách W_S.

- Bước 2 : Phân tích *T* (tách từ nếu cần thiết) để thu được danh sách W_T.

- Bước 3 : Với mỗi từ *s* trong W_s, ta chuyển *s* về dạng gốc (stemming) và xác định các ngữ nghĩa tương ứng trong từ điển song ngữ để xây dựng danh sách DT_s.

- Bước 4 : Từ những kết quả đã được chuẩn bị ở các bước trên, ta tính độ tương đồng hình vị (morpheme) của các từ *dt* trong DT_s đối với tất cả các từ *wt* trong W_T theo công thức tính hệ số Dice như sau :

$$Sim(dt, wt) = \frac{2 \times | dt \cap wt |}{| dt | + | wt |} \quad (1.1)$$

  Trong đó :

  - |*dt*| và |*wt*| : số hình vị trong từ của *dt* và *wt*.

- $|dt \cap wt|$: số hình vị giao nhau trong từ của *dt* và *wt*.

- Bước 5 : Đối với mỗi cặp từ ($s \in W_S$, $t \in W_T$) ta tính độ tương tự giữa chúng theo công thức :

$$DTSim(s, t) = \max_{t \in DT_s} Sim(s,t) \quad (1.2)$$

- Bước 6 : Với mỗi $s \in W_S$, chọn $t^* \in W_T$ sao cho $DTSim(s, t^*) = \max_{t \in W_T} DTSim(s,t)$ và $DTSim(s, t^*) > h_1$, với h_1 là ngưỡng định trước, và thêm liên kết (s, t^*) vào danh sách liên kết *CONN*.

- Bước 7 : Kết xuất danh sách liên kết *CONN*.

2.2.2.1.2 THUẬT TOÁN CLASS ALIGN

- Bước 1 : Chạy *DictAlign* trên tất cả các câu trong tập ngữ liệu song ngữ để lấy được danh sách những liên kết ứng viên *ALLCONN*.

- Bước 2 : Với mỗi lớp ngữ nghĩa $X \in CX$ (ngôn ngữ nguồn) và $Y \in CY$ (ngôn ngữ đích), ta tính độ tương đồng ngữ nghĩa *ClassSim(X, Y)* theo công thức sau :

$$ClassSim(X, Y) = \frac{\sum_{a \in X} From(a,Y) + \sum_{b \in Y} To(X,b)}{|X| + |Y|} \quad (1.3)$$

Trong đó :

- $|X|,|Y|$ lần lượt là tổng số từ trong lớp ngữ nghĩa *X*, *Y*.

- $From(a, Y) = 1$ nếu tồn tại $y \in Y$ sao cho $(a, y) \in ALLCONN$ và = 0 trong trường hợp ngược lại.

- *To(X, b) = 1* nếu tồn tại $x \in X$ sao cho $(x, b) \in ALLCONN$ và = 0 trong trường hợp ngược lại.

- *ALLCONN* là tập hợp tất cả những liên kết ứng viên có được khi chạy *DictAlign* trên toàn bộ các văn bản song ngữ.

- Bước 3 : Nếu *ClassSim(X, Y)* > h_2 (ngưỡng định trước) và *ClassSim(X, Y)* đạt cực đại với mọi *X* (cố định *Y*) hay cực đại với mọi *Y* (cố định *X*) thì ta thêm luật liên kết (*X*, *Y*) vào tập luật *RULES*.

- Bước 4 : Xuất kết quả là tập luật *RULES*.

2.2.2.1.3 THUẬT TOÁN CLASS BASED WORD ALIGNMENT

- Bước 1 : Phân tích câu *S* và câu *T* (tách từ, gán nhãn ngữ pháp, chuyển đổi về dạng gốc) để thu được W_S và W_T giống như trong thuật toán *DictAlign*.

- Bước 2 : Khởi tạo danh sách liên kết từ *ANN* rỗng. Chạy thuật toán *DictAlign* trên cặp câu (*S*, *T*) để thu được danh sách liên kết thô *CONN*.

- Bước 3 : Với mỗi ứng viên liên kết từ $(s, t) \in CONN$, ta tính xác suất liên kết *Pr(s, t)* theo công thức sau :

$$Pr(s, t) = T(s, t) \times D(i, j) \quad (1.4)$$

Trong đó :

- *T(s, t)* là xác suất dịch của hai từ *s* và *t*.
- *D(i, j)* là xác suất liên kết vị trí của hai từ ở vị trí *i* và *j* trong câu nguồn và câu đích.

- Bước 4 : Cập nhật liên kết cho danh sách *ANS* :

    - Bước 4.1 : Chọn từ *CONN* ứng viên liên kết (*s*, t**) thỏa điều kiện xác suất *Pr(s*, t*)* đạt cực đại so với mọi (*s, t*) và lớn hơn ngưỡng xác suất định trước h_3.

    - Bước 4.2 : Thêm (*s*, t**) vào *ANS* và loại bỏ khỏi *CONN*.

    - Bước 4.3 : Loại bỏ khỏi *CONN* những liên kết mâu thuẫn với (*s*, t**) và lặp lại bước 4.1 cho đến khi không thể chọn được ứng viên liên kết nào nữa.

- Bước 5 : Kết xuất danh sách *ANS* – kết quả cuối cùng của quá trình liên kết từ cho cặp câu (*S, T*).

Trong thuật toán *Class-based Word Alignment* mô tả ở trên :

- Xác suất dịch *T(s, t)* được tính như sau :

    - *T(s, t)* = t_1 nếu *ConceptSim(s, t)* $\geq h_1 \wedge DTSim(s,t) \geq h_2$
    - *T(s, t)* = t_2 nếu *ConceptSim(s, t)* $\geq h_1 \wedge DTSim(s,t) < h_2$
    - *T(s, t)* = t_3 nếu *ConceptSim(s, t)* $< h_1 \wedge DTSim(s,t) \geq h_2$
    - *T(s, t)* = t_4 nếu *ConceptSim(s, t)* $< h_1 \wedge DTSim(s,t) \geq h_2$

    Các hệ số t_1, t_2, t_3, t_4 được xác định theo nguyên tắc ước lượng cực đại (MLE – Maximum Likelihood Estimation). Theo đó, giả thiết có tổng cộng *n* mẫu cho liên kết từ (*s, t*) trên toàn bộ ngữ liệu và trong đó có *k* mẫu thỏa điều kiện *ConceptSim(s, t)* $\geq h_1 \wedge DTSim(s,t) \geq h_2$ thì khi đó $t_1 = k / n$. Tương tự như vậy, ta sẽ ước lượng được cho t_2, t_3 và t_4.

- Xác suất liên kết vị trí $D(i, j)$ được tính như sau :

  - $D(i, j) = d_1$ nếu $dist(i, j) = 0$.
  - $D(i, j) = d_2$ nếu $dist(i, j) = 1$.
  - $D(i, j) = d_3$ nếu $dist(i, j) = 2$.
  - $D(i, j) = d_4$ nếu $dist(i, j) \geq 3$.

  Trong đó, $dist(i, j) = |j - j'|$ nếu $\exists j' : (i, j') \in CONN$, còn trong trường hợp ngược lại thì $dist(i, j) = min(|(j - j_L) - (i - i_L)|, |(j - j_R) - (i - i_R)|)$. Với :

  - $(i_L, j_L) = \arg\max_{(i', j') \in CONN_{<i}} i'$
  - $(i_R, j_R) = \arg\min_{(i', j') \in CONN_{>i}} i'$

  Quan hệ $CONN_{<i}$ và $CONN_{>i}$ có thể hiểu là mối liên kết (i', j') với i' lớn nhất (nhỏ nhất) sao cho i' bé hơn (lớn hơn) i. Các hệ số d_1, d_2, d_3 và d_4 cũng được xác định bằng ước lượng cực đại (Maximum Likelihood Estimation) như đã mô tả ở trên.

2.2.2.2 LIÊN KẾT TỪ DỰA TRÊN XÁC SUẤT THỐNG KÊ

Brown et al. (1993) [7] đã đưa ra mô hình toán học cho dịch máy thống kê dựa trên cơ sở xác suất thống kê. Trong nghiên cứu đó, tác giả đã nghiên cứu và đề xuất ra năm mô hình thống kê nhằm mô phỏng quy trình dịch một câu từ ngôn ngữ nguồn sang ngôn ngữ đích (theo quan điểm xác suất thống kê). Trong số năm mô hình đó, ba mô hình đầu có liên quan chặt chẽ đến việc liên kết từ trong ngữ liệu song ngữ.

Về mặt bản chất, các mô hình toán học đó là nhằm giải thích theo quan điểm xác suất thống kê cách thức mà các từ trong ngôn ngữ nguồn liên kết với các từ trong ngôn ngữ đích. Trong phần này, chúng ta sẽ lần lượt khảo sát từng mô hình và cách áp dụng giải thuật ước lượng trung bình cực đại (Expectation – Maximization) để xác định các tham số của mô hình.

Khi nhìn nhận vấn đề theo quan điểm xác suất thống kê, chúng ta thấy là giữa một cặp câu ngẫu nhiên (*S, T*) thì có rất nhiều cách liên kết từ có thể xảy ra ngẫu nhiên. Do đó, khi ta cần chọn một cách liên kết từ thì đương nhiên ta phải chọn cách liên kết từ có nhiều khả năng xảy ra nhất – có xác suất xuất hiện cao nhất. Vì vậy, chúng ta cần phải quan tâm đến xác suất liên kết $P(A = a \mid S = s, T = t)$ gồm một bộ ba biến ngẫu nhiên (*S, A, T)*. Trong đó, *S* là câu nguồn, *T* là câu đích và *A* là sự liên kết ngẫu nhiên giữa chúng.

Trước hết, chúng ta định nghĩa tập hợp các liên kết từ có thể có giữa cặp câu (*s, t*) là *A(s, t)*. Theo đó, nếu $s = s_1 s_2 ... s_L$ và $t = t_1 t_2 ... t_M$ thì một mẫu liên kết từ ngẫu nhiên trong *A(s, t)* sẽ có dạng $a = a_1 a_2 ... a_M$ với $a_i \in [0, L]$. Ở đây, chúng ta đặt giả thiết hạn chế là một từ trong câu đích (ngôn ngữ đích) chỉ có thể được sinh ra bởi tối đa một từ trong câu nguồn (ngôn ngữ nguồn). Như vậy, một từ t_i trong câu đích sẽ được sinh ra bởi từ s_{a_i} (từ ở vị trí a_i) trong câu nguồn. Riêng đối với trường hợp $a_i = 0$ ta xem như t_i được sinh ra ngẫu nhiên nhằm mục đích làm cho câu đích trở nên kết dính hơn.

Không mất tính tổng quát, chúng ta có :

$$P(a \mid s, t) = \frac{P(a, t \mid s)}{P(t \mid s)} \quad (2.1)$$

Theo công thức Bayes, ta có :

$$P(t \mid s) = \sum_{a \in A(s,t)} P(a, t \mid s) \quad (2.2)$$

Ta lại đặt giả thiết rằng xác suất xuất hiện của a_i và v_i chỉ phụ thuộc vào những gì đã xuất hiện trước đó. Điều đó có nghĩa là xác suất phát sinh liên kết a_i chỉ phụ thuộc vào câu nguồn *s* ban đầu, a_1^{i-1} *(i – 1* liên kết trước đó) và t_1^{i-1}(*i – 1* từ trong câu đích đã được sinh ra trước đó). Tương tự, xác suất phát sinh từ t_i cũng sẽ phụ thuộc vào câu nguồn *s*, a_1^i (*i* liên kết đã sinh ra – tính cả liên kết a_i vừa mới sinh ra) và t_1^{i-1}(*i – 1* từ trong câu đích đã được sinh ra trước đó).

Như vậy , với (*s, a, t*) bất kỳ, ta có thể khai triển *P(a, t | s)* như sau :

$$P(a, t \mid s) = P(M \mid s) \prod_{i=1}^{M} P(a_i \mid a_1^{i-1}, t_1^{i-1}, s) P(t_i \mid a_1^i, t_1^{i-1}, s) \quad (2.3)$$

Trong công thức (2.3), *P(M | s)* là xác suất để câu đích có độ dài *M* từ, cho trước câu nguồn *s*. Để đơn giản hóa, ta đặt *P(M | s)* = ε (ε là một hằng số rất nhỏ so với 1). Khi đó, công thức (2.3) có thể được viết lại như sau :

$$P(a, t \mid s) = \varepsilon \prod_{i=1}^{M} P(a_i \mid a_1^{i-1}, t_1^{i-1}, s) P(t_i \mid a_1^i, t_1^{i-1}, s) \quad (2.4)$$

Từ công thức (2.4), Brown đã dẫn xuất ra năm mô hình liên kết từ thống kê dựa trên việc triển khai $P(a_i \mid a_1^{i-1}, t_1^{i-1}, s)$ và $P(t_i \mid a_1^i, t_1^{i-1}, s)$. Trong phần còn lại của chương này, chúng ta sẽ khảo sát 3 mô hình liên kết từ đầu tiên của Brown.

2.2.2.2.1 MÔ HÌNH 1 – XÁC SUẤT DỊCH TỪ

Đây là mô hình đơn giản nhất trong số các mô hình đề xuất bởi Brown. Mô hình này chỉ tập trung vào xác suất dịch từ của từng từ riêng biệt. Cụ thể, mô hình này giả thiết là xác suất sản sinh ra liên kết từ a_i là hoàn toàn độc lập ngẫu nhiên – nghĩa là a_i có thể mang bất kỳ giá trị nguyên nào trong đoạn [0, L] với xác suất bằng nhau, và xác suất sản sinh ra từ t_i chỉ phụ thuộc vào duy nhất từ liên kết với nó – tức là s_{a_i}. Như vậy, trong mô hình 1, ta có thể viết :

$$\prod_{i=1}^{M} P(a_i \mid a_1^{i-1}, t_1^{i-1}, s) = \frac{1}{|A(s,t)|} = \frac{1}{(L+1)^M} \quad (2.5)$$

$$\prod_{i=1}^{M} P(t_i \mid a_1^{i}, t_1^{i-1}, s) = \prod_{i=1}^{M} P(t_i \mid s_{a_i}) \quad (2.6)$$

Từ (2.5) và (2.6), ta có thể viết lại (2.4) như sau :

$$P(a,t \mid s) = \frac{\varepsilon}{(L+1)^M} \prod_{i=1}^{M} P(t_i \mid s_{a_i}) \quad (2.7)$$

Thế công thức (2.7) vào (2.2), ta thu được :

$$P(t \mid s) = \frac{\varepsilon}{(L+1)^M} \sum_{a_1=0}^{L} \ldots \sum_{a_M=0}^{L} \prod_{i=1}^{M} P(t_i \mid s_{a_i}) = \frac{\varepsilon}{(L+1)^M} \prod_{j=1}^{M} \sum_{i=0}^{L} P(t_j \mid s_i) \quad (2.8)$$

Dễ dàng nhận thấy là công thức (2.8) hoàn toàn độc lập với *a*. Từ đó, theo công thức (1), để cực đại hóa *P(a | s, t)* thì ta chỉ cần cực đại hóa *P(a, t | s)* :

$$a^* = \arg\max_{a \in A(s,t)} \{ \frac{\varepsilon}{(L+1)^M} \prod_{i=1}^{M} P(t_i \mid s_{a_i}) \} \quad (2.9)$$

Trong mô hình này, các xác suất dịch $P(t_j \mid s_i)$ chính là các tham số cần ước lượng và để ước lượng những tham số này chúng ta cần dựa vào ngữ liệu song ngữ sẵn có.

Trước khi mô tả giải thuật ước lượng bộ tham số $P(t_j \mid s_i)$ cho mô hình nói trên, chúng ta sẽ thống nhất trên một số khái niệm sau :

- C – Ngữ liệu song ngữ ($|C| = S$).

- (s^r, t^r) – Cặp câu thứ r trong ngữ liệu song ngữ.

- $count(t_i \mid s_j; C)$ – Số lần từ t_i được sinh ra từ (liên kết với) s_j quan sát trên toàn ngữ liệu song ngữ .

- $count(t_i \mid s_j; s^r, t^r)$ – Số lần từ t_i được sinh ra từ (liên kết với) s_j quan sát trên cặp câu thứ r của ngữ liệu song ngữ.

- $\delta(x, y) = \begin{cases} 0 \Leftrightarrow x \neq y \\ 1 \Leftrightarrow x = y \end{cases}$ – Hàm biệt số Kronecker.

Không mất tính tổng quát, ta có thể viết :

$$P(t_i \mid s_j) = \frac{count(t_i \mid s_j; C)}{\sum_{t_r} count(t_r \mid s_j; C)} \quad (2.10)$$

$$count(t_i \mid s_j; C) = \sum_{r=1}^{S} count(t_i \mid s_j; t^r, s^r) \quad (2.11)$$

Công thức (2.10) và (2.11) chính là thể hiện cách ước lượng trực tiếp $P(t_i \mid s_j)$ theo nguyên tắc ước lượng cực đại (Maximum Likelihood Estimation). Tuy nhiên, không may là các tham số $count(t_i \mid s_j; t^r, s^r)$ không thể ước lượng trực tiếp bằng cách quan sát từng cặp câu trong ngữ liệu song ngữ vì bản thân ngữ liệu song ngữ của chúng ta không chứa thông tin về liên kết từ - là thông tin mà chúng ta đang tìm cách xây

dựng. Chính vì không thể đưa ra một tính toán chính xác cho $count(t_i \mid s_j; t^r, s^r)$ nên chúng ta buộc phải sử dụng ước lượng trung bình của $count(t_i \mid s_j; t^r, s^r)$ để thay thế :

$$count(t_i \mid s_j; t^r, s^r) = \sum_{a_1=0}^{L} \sum_{a_2=0}^{L} \ldots \sum_{a_M=0}^{L} P(a \mid t^r, s^r) \sum_{l=1}^{M} \delta(t_i, t_l^r) \delta(s_j, s_{a_l}^r) \quad (2.12)$$

Như thế, việc xác định liên kết từ và ước lượng tham số xem như tạo thành một vòng luẩn quẩn : muốn liên kết từ ta cần ước lượng tham số, nhưng để ước lượng tham số ta lại cần liên kết từ !

Để phá vỡ vòng luẩn quẩn đó, trong nghiên cứu của mình, Brown đã áp dụng giải thuật ước lượng trung bình cực đại (EM - Expectation Maximization) :

- B1 : Ban đầu, các tham số $P(t_i \mid s_j)$ của mô hình sẽ được khởi tạo ngẫu nhiên.
- B2 : Dựa trên các giá trị khởi tạo đó, các xác suất liên kết từ $P(a \mid s, t)$ sẽ được ước lượng theo công thức (2.7).
- B3 : Các tham số $P(t_i \mid s_j)$ sẽ được ước lượng lại dựa trên $P(a \mid s, t)$.
- B4 : Lặp lại B1 cho đến khi ngưỡng dao động của $P(t_i \mid s_j)$ đủ nhỏ.

Giải thuật EM nói trên, đã được chứng minh về cả lý thuyết và thực nghiệm là hệ sẽ hội tụ sau một số hữu hạn bước. Thật ra, mô tả giải thuật ở trên chỉ là mô tả hình thức về mặt lý thuyết chứ không khả thi về mặt tính toán vì rõ ràng là ta không thể ước lượng $P(a \mid s, t)$ cho mọi liên kết từ a vì không gian $A(s, t)$ có kích thước quá lớn. Như vậy, để có thể cài đặt được giải thuật một cách hiệu quả, ta cần một cách tính khéo léo hơn. Cụ thể, trong công thức (2.12), ta có thể thế công thức (2.1) vào và thu được :

$$count(t_i \mid s_j; t^r, s^r) = \sum_{a \in A(s,t)} \frac{P(a, t^r \mid s^r)}{P(t^r \mid s^r)} \sum_{l=1}^{M} \delta(t_i, t_l^r) \delta(s_j, s_{a_l}^r) \quad (2.13)$$

Thế các công thức (2.7), (2.8) vào (2.13) và triển khai, ta thu được :

$$count(t_i \mid s_j; t^r, s^r) = \frac{P(t_i \mid s_j)}{\sum_{l=0}^{L} P(t_i \mid s_l^r)} \sum_{u=1}^{M} \delta(t_i, t_u^r) \sum_{v=0}^{L} \delta(s_j, s_v^r) \quad (2.14)$$

Với công thức (2.14), ta đã loại bỏ hoàn toàn *a* trong biểu thức vế phải và điều đó giúp việc tính toán được cải thiện rất nhiều về tốc độ. Khi đó, giải thuật EM ở trên được mô tả lại chi tiết (về mặt ký thuật) như sau :

- B1 : Khởi tạo giá trị ngẫu nhiên cho $P(t_i \mid s_j)$ – khởi tạo ngẫu nhiên hay khởi tạo dạng xác suất phân bố đều.

- B2 : Với mỗi cặp câu $(s^r, t^r) \in C$, ta xác định các tham số $count(t_i \mid s_j ; t^r, s^r)$ theo công thức (2.14).

- B3 : Với mỗi cặp từ s_j, t_i có mặt trong ngữ liệu song ngữ *C*, ta ước tính lại tham số $P(t_i \mid s_j)$ theo công thức (2.10) và (2.11).

- B4 : Với bộ tham số $P(t_i \mid s_j)$ mới ước lượng lại ở B3, ta lặp lại B2 và quá trình cứ tiếp diễn đến khi nào ngưỡng dao động của $P(t_i \mid s_j)$ đủ nhỏ (hội tụ).

Khi đã ước lượng được bộ tham số xác suất dịch từ, ta có thể dễ dàng xác định liên kết từ tối ưu *a** trong công thức (2.9) như sau :

$$\forall 1 \le i \le M, a_i^* = \arg\max_j P(t_i \mid s_j) \quad (2.15)$$

2.2.2.2.2 MÔ HÌNH 2 – XÁC SUẤT DỊCH TỪ, LIÊN KẾT TỪ CỤC BỘ

Điểm hạn chế của cơ chế liên kết từ trong mô hình 1, như ta thấy trong công thức (2.15), là không tính đến yếu tố về trật từ từ - vị trí tương đối giữa các từ trong câu. Nói cho đơn giản là trong mô hình 1, nếu không kể đến các tham số $P(t_i \mid s_j)$ thì một từ của câu nguồn cũng có thể liên kết với bất kỳ từ nào ở câu đích mà không bị ràng buộc, hạn chế nào ngay cả khi một từ đứng đầu câu, một từ đứng cuối câu ! Nguyên nhân chính của khuyết điểm đó là do ngay từ đầu, mô hình 1 đã đặt giả thiết là $P(a_i \mid a_1^{i-1}, t_1^{i-1}, s)$ độc lập ngẫu nhiên, theo công thức (2.5). Để khắc phục hạn chế đó, mô hình 2 đã đưa vào xem xét giả thiết rằng $P(a_i \mid a_1^{i-1}, t_1^{i-1}, s)$ phụ thuộc vào vị trí I của từ đang xét trong câu đích, độ dài câu đích m và độ dài câu nguồn l :

$$P(a_i \mid a_1^{i-1}, t_1^{i-1}, s) = \alpha(a_i = j \mid i, m, l) = \alpha(j \mid i, m, l) \quad (2.16)$$

$$\sum_{j=0}^{l} \alpha(j \mid i, m, l) \backslash = 1 \forall (i, m, l) \quad (2.17)$$

Từ (2.16), (2.17) ta có thể viết lại (2.7) và (2.8) như sau :

$$P(a, t \mid s) = \frac{\varepsilon}{(l+1)^m} \prod_{j=1}^{m} P(t_j \mid s_{a_j}) \alpha(a_j \mid j, m, l) \quad (2.18)$$

$$P(t \mid s) = \frac{\varepsilon}{(l+1)^m} \sum_{a_1=0}^{l} \ldots \sum_{a_m=0}^{l} \prod_{j=1}^{m} P(t_j \mid s_{a_j}) \alpha(a_j \mid j, m, l) \quad (2.19)$$

Về bản chất, việc ước lượng tham số và tính liên kết từ tối ưu a^* trong mô hình 2 căn bản là giống mô hình 1 chỉ khác là để ước lượng các tham số $\alpha(i \mid j, m, l)$ thì ta cần phải khảo sát thêm *count(i | j; m, l, t^r, s^r)*. Cũng giống như *count(t_i | s_j ; t^r, s^r)*, *count(i | j; m, l, t^r, s^r)* không thể tính trực tiếp do ngữ liệu song ngữ không chứa thông tin về xác suất liên kết từ, là thông tin mà chúng ta đang cố gắng xây dựng, và điều đó buộc chúng ta phải sử dụng tạm ước lượng trung bình của *count(i | j; m, l, t^r, s^r)*. Như thế, không mất tính tổng quát ta có :

$$count(i \mid j; m, l, t^t, s^r) = \sum_{a \in A(s,t)} P(a \mid s, t)\delta(i, a_j) \quad (2.20)$$

$$\alpha(i \mid j, m, l) = \frac{\sum_{s=1}^{S} count(i \mid j; m, l, t^r, s^r)}{\sum_{k}\sum_{s=1}^{S} count(k \mid j; m, l, t^r, s^r)} \quad (2.21)$$

Biến đổi và triển khai tương tự như ở mô hình 1, ta thu được :

$$count(t_i \mid s_j; t^r, s^r) = \sum_{j'=1}^{m}\sum_{i'=0}^{l} \frac{P(t_i \mid s_j)\alpha(j \mid i, m, l)\delta(t_i, t_{i'})\delta(s_j, s_{j'})}{P(t_i \mid s_0)\alpha(0 \mid i, m, l) + ... + P(t_i \mid s_l)\alpha(l \mid i, m, l)} \quad (2.22)$$

$$count(i \mid j; m, l, t^r, s^r) = \frac{P(t_i \mid s_j)\alpha(j \mid i, m, l)}{P(t_i \mid s_0)\alpha(0 \mid i, m, l) + ... + P(t_i \mid s_l)\alpha(l \mid i, m, l)} \quad (2.23)$$

Với hai công thức (2.22), (2.23) vừa thu được ở trên, ta có nhận xét là về bản chất thì mô hình 1 chính là trường hợp đặc biệt của mô hình 2 khi ta giả định rằng $\alpha(i \mid j, m, l) = \frac{1}{l+1}$. Nói chung, các bước thực hiện ước lượng cực đại cho các tham số của mô hình 2 cũng tương tự như mô hình 1, ngoại trừ việc các tham số $P(t_i \mid s_j)$ thay vì được khởi tạo ngẫu nhiên thì sẽ lấy giá trị đã hội tụ từ mô hình 1. Mục đích của việc chuyển giao thông tin từ mô hình trước sang mô hình sau như vậy chính là nhằm cải thiện chất lượng và tiết kiệm thời gian tính toán phức tạp.

2.2.2.2.3 MÔ HÌNH 3 – XÁC SUẤT GIÁ TRỊ SẢN SINH

Mô hình 3 xuất phát từ một góc nhìn khác hẳn so với mô hình 1 và 2 về quá trình dịch thống kê. Như chúng ta đã khảo sát, mô hình 1 và 2 về bản chất xem quá trình dịch thống kê như một quy trình ngẫu nhiên tuần tự - công thức (2.4):

- Đầu tiên, từ câu nguồn *s* với độ dài *l* từ, độ dài *m* của câu đích được sinh ra ngẫu nhiên .

- Sau đó, lần lượt từng vị trí của của câu đích được xem xét, và sinh ra ngẫu nhiên các liên kết từ tương ứng a_i.

- Từ các liên kết tương ứng trong câu nguồn, các từ trong câu đích *t* lần lượt được sinh ra ngẫu nhiên theo xác suất dịch từ.

Đối với mô hình 3, quy trình ngẫu nhiên trong dịch thống kê lại được quan sát ở một khía cạnh khác. Trong mô hình này, Brown đưa vào khái niệm giá trị sản sinh của một từ bất kỳ trong ngôn ngữ nguồn như một biến ngẫu nhiên ϕ tham gia vào quy trình ngẫu nhiên của dịch thống kê. Quy trình ngẫu nhiên được giả định trong mô hình 3 có thể được mô tả ngắn gọn như sau :

- Ban đầu, giá trị sản sinh ϕ_i của mỗi từ s_i trong câu nguồn *s* được sinh ra ngẫu nhiên.

- Sau khi giá trị sản sinh được sinh ra ngẫu nhiên, quy trình sẽ bắt đầu sinh ra ngẫu nhiên các từ trong câu đích (ở dạng túi từ τ - chưa xét đến vị trí tương đối trong câu).

- Sau khi các từ trong câu đích đã được sinh ra, một hoán vị π trên vị trí của chúng trong câu sẽ được sinh ra ngẫu nhiên.

Như thế, rõ ràng về bản chất từ (τ, π) ta có thể suy ra (*a*, *t*) và vì thế, không mất tính tổng quát, ta có công thức sau :

$$P(a,t \mid s) = \sum_{(\tau,\pi)\in<a,t>} P(\tau,\pi \mid s) \quad (2.24)$$

Phỏng theo quy trình ngẫu nhiên ở trên, $P(\tau, \pi \mid s)$ sẽ được khai triển tuần tự như sau :

$$P(\tau, \pi \mid s) = P(\phi \mid s)P(\tau \mid \phi, s)P(\pi \mid \phi, \tau, s) \quad (2.25)$$

$$P(\phi \mid s) = P(\phi_0 \mid \phi_1^l, s)\prod_{i=1}^{l} P(\phi_i \mid \phi_1^{i-1}, s) \quad (2.26)$$

$$P(\tau \mid \phi, s) = \prod_{i=0}^{l}\prod_{k=1}^{\phi_i} P(\tau_{i,k} \mid \tau_{i,1}^{k-1}, \tau_0^{i-1}, \phi_0^l, s) \quad (2.27)$$

$$P(\pi \mid \phi, \tau, s) = \prod_{i=1}^{l}\prod_{k=1}^{\phi_i} P(\pi_{i,k} \mid \pi_{i,1}^{k-1}, \pi_1^{i-1}, \tau_0^l, \phi_0^l, s)\prod_{k=1}^{\phi_0} P(\pi_{0,k} \mid \pi_{0,1}^{k-1}, \pi_1^l, \tau_0^l, \phi_0^l, s) \quad (2.28)$$

Trong các công thức (2.25), (2.26), (2.27) và (2.28), ta chú giải cho các ký hiệu sử dụng như sau :

- ϕ_i giá trị sản sinh của từ thứ *i* trong câu nguồn *s*; trong đó, ϕ_0 là giá trị sản sinh của phần tử rỗng s_0 = *NULL* và ngoài ra, ta ký hiệu ϕ_1^l là tập hợp các giá trị sản sinh của các từ $s_1, s_2, \ldots, s_l$.

- τ_i là nhóm từ (túi từ) trong ngôn ngữ đích được sinh ra bởi từ s_i, ngoài ra, ta còn ký hiệu τ_1^l là tập hợp các túi từ sinh ra bởi các từ $s_1, s_2, \ldots, s_l$ và $\tau_{i,1}^k$ là tập hợp *k* từ đầu tiên trong túi từ τ_i.

- π_i là tập hợp các vị trí tương ứng với các từ trong túi từ τ_i trên câu đích, thêm nữa, ta cũng ký hiệu π_1^l là tập hợp các $\pi_1, \pi_2, \ldots, \pi_l$ và $\pi_{i,1}^k$ là tập hợp *k* phần tử đầu trong π_i.

Từ các công thức (2.25), (2.26), (2.27) và (2.28) mô tả ở trên, Brown đã tạm đặt ra các giả thiết đơn giản hóa để hiện thực hóa mô hình 2. Theo đó, Brown giả định

rằng $P(\phi_i \mid \phi_1^{i-1}, s)\ (1 \le i \le l)$ chỉ phụ thuộc vào s_i - $P(\phi_i \mid \phi_1^{i-1}, s) = n(\phi_i \mid s_i)$, $P(\tau_{i,k} \mid \tau_{i,1}^{k-1}, \tau_0^{i-1}, \phi_0^l, s)$ chỉ phụ thuộc vào s_i - $P(\tau_{i,k} \mid \tau_{i,1}^{k-1}, \tau_0^{i-1}, \phi_0^l, s) = P(t_j \mid s_i)$ và $P(\pi_{i,k} \mid \pi_{i,1}^{k-1}, \pi_1^{i-1}, \tau_0^l, \phi_0^l, s)\ (1 \le i \le l)$ chỉ phụ thuộc vào *i, m, l* - $P(\pi_{i,k} \mid \pi_{i,1}^{k-1}, \pi_1^{i-1}, \tau_0^l, \phi_0^l, s)$ = *P(j | i, m, l)*. Các xác suất giá trị sản sinh $P(\phi_0 \mid \phi_1^l, s)$ và xác suất hoán vị $P(\pi_{0,k} \mid \pi_{0,1}^{k-1}, \pi_1^l, \tau_0^l, \phi_0^l, s)$ được xét riêng. Theo Brown, $P(\pi_{0,k} \mid \pi_{0,1}^{k-1}, \pi_1^l, \tau_0^l, \phi_0^l, s)$ là xác suất sinh ra vị trí các từ trong τ_0, là những từ được xem xét cách đặt cuối cùng, nên khi sinh ra $\pi_{0,k}$ chúng ta chỉ còn lại $\phi_0 - k$ vị trí để xét, vì vậy, $P(\pi_{0,k} \mid \pi_{0,1}^{k-1}, \pi_1^l, \tau_0^l, \phi_0^l, s) = \frac{1}{\phi_0 - k}$. Từ đó, chúng ta có thể viết :

$$\prod_{k=1}^{\phi_0} P(\pi_{0,k} \mid \pi_{0,1}^{k-1}, \pi_1^l, \tau_0^l, \phi_0^l, s) = \frac{1}{\phi_0!} \quad (2.29)$$

Ngoài ra, chúng ta cũng có $P(\phi_0 \mid \phi_1^l, s)$ là xác suất sinh ra được ϕ_0 từ trong câu đích bằng s_0 khi đã sinh ra $\sum_{i=1}^{l} \phi_i$ từ bằng s_i :

$$P(\phi_0 \mid \phi_1^l, s) = \binom{\phi_1 + ... + \phi_l}{\phi_0} p_0^{\phi_1 + ... + \phi_l - \phi_0} p_1^{\phi_0} \quad (2.30)$$

Trong công thức (2.29), p_0 là xác suất một từ bất kỳ trong câu đích được sinh ra từ phần từ $s_0 = NULL$ và $p_1 = 1 - p_0$ là xác suất một từ bất kỳ trong câu đích được sinh ra từ s_i. Công thức (2.29) có thể hiểu là xác suất để có ϕ_0 từ trong tổng số $m = \sum_{i=0}^{l} \phi_i$ sinh ra từ s_0, cho trước xác suất để sinh một từ bất kỳ từ s_0 là p_0 (theo định luật phân bố xác suất nhị thức). Như vậy, sau khi đã đưa ra một loạt giả thiết nhằm đơn giản hóa các tham số thì ta có thể viết lại công thức của mô hình 3 cụ thể như sau :

$$P(a,t \mid s) = \binom{m-\phi_0}{\phi_0} p_0^{m-2\phi_0} p_1^{\phi_0} \prod_{i=1}^{l} \phi_i! P(\phi_i \mid s_i) \prod_{i=1}^{m} P(t_i \mid s_{a_i}) P(i \mid a_i, m, l) \quad (2.31)$$

Trong công thức (2.31), các điều kiện liên quan đến các tham số sau đây phải được thỏa mãn : $\sum_{t_i} P(t_i \mid s_j) = 1$, $\sum_{j} P(j \mid i,m,l) = 1$, $\sum_{\phi_i} P(\phi_i \mid s_i) = 1$, $p_0 + p_1 = 1$. Ngoài ra, để ước lượng cực đại cho các tham số nói trên trong mrô hình 3, ta cũng làm giống như mô hình 1 và 2.

2.2.2.3 KẾT LUẬN, ĐÁNH GIÁ

Trong các phần trước, chúng ta đã khảo sát hai mô hình liên kết từ phổ biến nhất hiện nay. Những mô hình này đều có những ưu / khuyết điểm riêng nên việc sử dụng mô hình nào trong thực nghiệm phụ thuộc vào bản chất vấn đề cùng các nguồn tài nguyên hỗ trợ. Nhìn chung, liên kết từ theo mô hình Semantic Class (SC) [13] thường cho độ chính xác cao nhưng độ phủ lại kém còn liên kết từ theo mô hình Statistical Machine Translation (SMT) [7] thì độ phủ cao nhưng lại có độ chính xác thấp. Gần đây, để nâng cao chất lượng liên kết từ bằng thống kê, người ta đang cố gắng tích hợp các nguồn thông tin bổ sung về ngôn ngữ như ranh giới từ, từ loại, … vào các mô hình thống kê.

CHƯƠNG 3 – MÔ HÌNH KHỬ NHẬP NHẰNG NGỮ NGHĨA

3.1 CÁC MÔ HÌNH XỬ LÝ NHẬP NHẰNG NGỮ NGHĨA ĐÃ SỬ DỤNG

Các phương pháp khử nhập nhằng ngữ nghĩa thường được phân loại dựa trên cách thức tiếp cận bản chất của ngôn ngữ của chúng là có giám sát hay không có giám sát (supervised or unsupervised learning). Với cách tiếp cận có giám sát, chúng ta cần một kho ngữ liệu được gán nhãn ngữ nghĩa hoàn chỉnh, từ đó tiến hành học mẫu để nhận biết và phân loại. Với cách tiếp cận không giám sát, quy trình học có thể được hình dung như một quá trình gom nhóm các mẫu học chưa được gán nhãn để từ đó rút ra tri thức.

Một cách phân loại khác là dựa trên bản chất tài nguyên sử dụng trong quá trình học máy là hướng ngữ liệu hay hướng từ điển (corpus-based or dictionary-based). Với các cách tiếp cận hướng từ điển, tài nguyên thường được sử dụng là các thể học (ontology) như MRD (machine readable dictionary), WordNet. Nguyên tắc của các cách tiếp cận này nói chung là dựa trên mối liên hệ ngữ nghĩa (synonym, hypernym, hyponym, …) giữa các từ để xây dựng bộ luật hướng ngữ cảnh.

3.1.1 KHỬ NHẬP NHẰNG NGỮ NGHĨA HƯỚNG TỪ ĐIỂN

3.1.1.1 SỬ DỤNG ĐỊNH NGHĨA TRONG TỪ ĐIỂN

Nền tảng của phương pháp khử nhập nhằng ngữ nghĩa hướng từ điển là dựa trên ý tưởng của Lesk (1986) [1] rằng những từ *v* nằm trong định nghĩa của từ đa nghĩa *w* trong từ điển rất có khả năng là những từ chỉ định cho sắc thái ý nghĩa tương ứng mà chúng xác định.

Ví dụ : Trong từ điển từ *cone* có hai sắc thái ý nghĩa sau đây :

- Bracts in trees of pine family.
- A crisp of cone-shaped wafer for holding ice cream.

Như vậy, nếu trong ngữ cảnh của từ *cone* có sự xuất hiện của *tree* hay *ice* thì nhiều khả năng từ *cone* sẽ mang sắc thái ý nghĩa tương ứng. Ý tưởng đó được Lesk tổng quát hóa lên thành giải pháp như sau :

- Gọi $D_1, D_2, ..., D_K$ là các định nghĩa của các sắc thái ý nghĩa $s_1, s_2, ..., s_K$ trong từ điển của từ đa nghĩa *w*. (D_i được xem như một túi từ (bag of words))
- Gọi E_v là định nghĩa trong từ điển của từ *v* xuất hiện trong ngữ cảnh *c* của *w*. Nếu *v* có nhiều sắc thái ý nghĩa thì xem như E_v là tập hợp tất cả các từ xuất hiện trong các định nghĩa của *v*.
- Tính $score(s_i) = overlap(D_i, \bigcup_{v \in c} E_v)$
- Sắc thái ý nghĩa tương ứng với ngữ cảnh *c* của từ đa nghĩa *w* sẽ là $s' = argmax_s\{score(s)\}$.

Hàm *overlap* nói trên có thể cài đặt đơn giản hay phức tạp nhưng nói chung là nhằm phản ánh mức độ liên quan về mặt ngữ nghĩa của ngữ cảnh đối với sắc thái ý nghĩa. Một cách cài đặt đơn giản cho hàm *overlap* là đếm số từ chung (sau khi đưa về dạng gốc – bỏ qua hình thái) giữa hai túi từ. Tuy nhiên, thực nghiệm cho thấy hiệu quả của cách làm này không cao. Tỉ lệ xác định đúng sắc thái ý nghĩa của các từ đa nghĩa chỉ dao động trong khoảng từ 50% đến 70%.

3.1.1.2 SỬ DỤNG PHẠM TRÙ NGỮ NGHĨA

Nhằm mục đích cải tiến, Yarowsky (1992) [4] đã đề xuất một phương pháp khử nhập nhằng ngữ nghĩa hướng từ điển kết hợp với ngữ liệu đơn ngữ với độ chính xác khá cao. Ý tưởng của Yarowsky cho rằng bản chất sắc thái ý nghĩa của các từ đa nghĩa gắn liền với các phạm trù (category) trong ngôn ngữ. Nói cách khác, sắc thái

ý nghĩa của một từ có thể xem như một phạm trù ngôn ngữ mà từ đó thuộc về. Từ đó, vấn đề khử nhập nhằng ngữ nghĩa được quy về việc xét xem khi đặt trong ngữ cảnh *c* thì từ đa nghĩa *w* sẽ thuộc về phạm trù nào. Tất nhiên, để làm được điều đó thì từ điển dùng làm tài nguyên phải là từ điển lớn có sự phân chia, sắp xếp các từ theo các nhóm phạm trù ngôn ngữ như Roget (Roget, 1946) hay Longman (Procter, 1978).

Nguyên tắc cơ bản trong cách tiếp cận của Yarowsky là dựa vào các phạm trù của các từ trong ngữ cảnh (được định nghĩa trong từ điển) để xác định phạm trù của ngữ cảnh đó và cuối cùng là dựa vào đó để khử nhập nhằng ngữ nghĩa. Tuy nhiên, có một vấn đề phát sinh từ một thực tế là cho dù từ điển dùng làm tài nguyên có lớn đến đâu thì cũng khó mà bao trùm được tất cả các từ cùng với phạm trù của chúng. Ngoài ra, những từ điển lớn như vậy thường là rất tổng quát do đó phạm trù gán cho các từ cũng ở mức tổng quát và chính vì thế, đôi khi điều này sẽ làm sai lệch kết quả dự đoán phạm trù ngôn ngữ khi ta tiến hành phân tích trên những văn bản chuyên về một lĩnh vực nào đấy.

Ví dụ :

- Using the mouse, I could draw various types of animals such as bat, tiger, shark, bird, lion, dolphin and elephant on the screen of my computer. (1)
- Nadal beats Federer in Wimbledon 2008. (2)

Ở ví dụ thứ nhất, từ *mouse* rất dễ bị gán cho sắc thái ý nghĩa chỉ loài chuột, một loài động vật do những từ trong ngữ cảnh như *animals, bat, tiger, shark, bird, elephant, lion, dolphin* trong từ điển hầu hết thuộc về phạm trù ngữ nghĩa động vật, trong khi những từ chỉ đến phạm trù thiết bị máy tính như *screen, computer* thì lại quá ít. Đó chính là vấn đề về mức tổng quát của từ điển nảy sinh khi phân tích những văn bản thuộc về một lĩnh vực nhỏ (ở đây là về thiết bị máy tính). Ở ví dụ thứ hai, từ *beats* có khả năng bị hiểu nhầm là hành động đánh chứ không phải là sự chiến thắng trong

một trận đấu quần vợt. Lý do là ngữ cảnh của nó quá nghèo nàn (những từ như *Nadal, Federer, Wimbledon* không hề có trong từ điển, mặc dù đó mới chính là những từ đóng góp chủ yếu vào ngữ cảnh của từ *beat*).

Rõ ràng, để giải quyết vấn đề trên cách duy nhất là sử dụng một nguồn ngữ liệu chuyên về lĩnh vực đang xử lý để có thêm thông tin hỗ trợ. Đó là lý do tại sao trong cách tiếp cận của Yarowsky cần có sự kết hợp của ngữ liệu đơn ngữ. Yarowsky đã tổng quát hóa ý tưởng trên thành phương pháp tiếp cận như sau :

- B1 - Xem như ngữ liệu đơn ngữ của ta là tập các ngữ cảnh *C* của từ đa nghĩa *w*.
- B2 - Với mỗi phạm trù ngôn ngữ (có thể xem như sắc thái ý nghĩa) *t* được liệt kê trong từ điển, và với mỗi ngữ cảnh *c* xuất hiện trong ngữ liệu ta khởi tạo theo nguyên lý Greedy giá trị $score(c,t) = \sum_{v \in c} \delta(t,v)$ với $\delta(t,v) = 1$ nếu *t* là một trong những phạm trù ngữ nghĩa của *v* và $\delta(t,v) = 0$ trong trường hợp ngược lại.
- B3 - Với mỗi ngữ cảnh c_i, ta xây dựng tập các phạm trù ứng viên của c_i như sau: $T(c_i) = \{t \mid score(c_i, t) > threshold\}$. Trong đó, *threshold* là một ngưỡng đủ lớn được định trước để loại bỏ đi những phạm trù ít khả năng là chủ đề của ngữ cảnh nói trên.
- B4 - Với mỗi từ *v* trong từ điển, ta xây dựng tập $V = \{c \mid v \in c\}$.
- B5 - Với mỗi phạm trù (chủ đề) *t*, ta xây dựng tập $T = \{c \mid t \in T(c)\}$.
- B6 - Tính các tham số của mô hình :

$$P(v_j \mid t_i) = |V_j \cap T_i| / \sum_j |V_j \cap T_i| \quad (3.1)$$

$$P(t_i) = (\sum_j |V_j \cap T_i|) / (\sum_i \sum_j |V_j \cap T_i|) \quad (3.2)$$

- B7 - Ước lượng lại thông số *score(.)* theo các tham số mới tính :

$$score(c,t) = \ln \frac{P(c \mid t)}{P(c)} P(t) \quad (3.3)$$

$$P(c \mid t) = \prod_{v \in c} P(v \mid t) \quad (3.4)$$

$$P(c) = \prod_{v \in c} P(v) \quad (3.5)$$

- B8 - Lặp lại bước B3 chừng nào khoảng dao động của các tham số mô hình còn lớn hơn một ngưỡng hội tụ định sẵn.
- B9 - Với mỗi phạm trù (sắc thái ý nghĩa) t_k của từ đa nghĩa *w* ta tính xác suất ứng với ngữ cảnh *c* như sau :

$$score(t_k) = \ln P(t_k) + \sum_{v \in c} \ln P(v \mid t_k) \quad (3.6)$$

- B10 - Cuối cùng, ta quyết định sắc thái ý nghĩa ứng với ngữ cảnh *c* của từ đa nghĩa *w* là $t = argmax_t \{score(t)\}$.

Phương pháp trên thu được hiệu quả khá tốt khi thử nghiệm nhưng vẫn có một điểm yếu ở chỗ nó không xử lý được những trường hợp mà một sắc thái ý nghĩa có thể xuất hiện trong nhiều phạm trù ngôn ngữ (trái với giả thiết của Yarowsky). Ví dụ như sắc thái ý nghĩa của từ *interest* trong *self-interest* có thể xuất hiện trong nhiều phạm trù như âm nhạc (*music*), giải trí (*entertainment*), tài chính (*finance*). Thực nghiệm cho thấy là trong những trường hợp như thế thì giải thuật thường cho kết quả không tốt.

3.1.1.3 ONE SENSE PER DISCOURSE

Ý tưởng của phương pháp tiếp cận này dựa trên hai nhận xét sau đây của Yarowsky (1995) [9]:

- Nếu một từ đa nghĩa *w* xuất hiện nhiều lần trong cùng một văn bản hay một đoạn văn thì gần như tất cả các lần xuất hiện đó đều có cùng một sắc thái ý nghĩa. (One sense per discourse)
- Những từ hay cụm từ xuất hiện trong cửa sổ ngữ cảnh xung quanh từ đa nghĩa *w* cung cấp thông tin vững chắc và nhất quán về sắc thái ý nghĩa của *w*. (One sense per collocation)

Nhận xét thứ nhất của Yarowsky nhằm khắc phục một khuyết điểm trong hai cách tiếp cận nói trên là xem xét những lần xuất hiện khác nhau của một từ đa nghĩa *w* một cách riêng biệt. Trên thực tế, nếu những lần xuất hiện đó ở trong cùng một văn bản thì lẽ tất nhiên chúng phải có sắc thái ý nghĩa nhất quán với nhau. Nói cách khác là có một sự ràng buộc về tính nhất quán về mặt ngữ nghĩa trong phạm vi một văn bản.

Ví dụ : Xét văn bản sau :

… the existence of plant and animal life … classified as either plant or animal … Although bacterial and plant cells are enclosed …

Trong văn bản trên từ *plant* xuất hiện ba lần. Ở hai lần xuất hiện đầu, trong ngữ cảnh của *plant* có cụm từ *animal, life* nên có thể dễ dàng phân biệt *plant* ở đây là chỉ dạng sống thực vật. Tuy nhiên, ở lần xuất hiện thứ ba thì ngữ cảnh tương đối nghèo nàn hơn do vậy rất có khả năng từ *plant* bị phân biệt sai là một dạng máy móc thiết bị. Trong tình huống như vậy, nếu ta xét đến tính toàn vẹn nhất quán về mặt ngữ nghĩa của từ *plant* trong cùng một văn bản thì rõ ràng *plant* phải được phân biệt là một dạng sống thực vật (*living being*) do đó là sắc thái ý nghĩa xuất hiện nhiều nhất của nó trên toàn văn bản.

Nhận xét thứ hai của Yarowsky thật ra cũng chính là nhận xét cơ sở trong lĩnh vực xử lý nhập nhằng ngữ nghĩa. Tuy nhiên, Yarowsky đã phát triển nó ở một mức tinh vi hơn. Nói một cách nào đấy thì nhận xét này cũng tương tự như cách tiếp cận của Brown et al. (1991) [2] trong mô hình Information Theory. Cụ thể, Yarowsky cho rằng cách tốt nhất để phân biệt ngữ nghĩa chính xác trong mô hình này là dựa trên những cụm từ đồng hiện (collocational words / phrases). Theo đó, những cụm từ đồng hiện f thường xuất hiện với một sắc thái ý nghĩa s nào đó sẽ được tìm kiếm và lưu trữ trong tập đặc trưng của s. Tập đặc trưng này sau đó sẽ được sử dụng cho

việc xác định xem sắc thái ý nghĩa nào của từ đa nghĩa *w* là trội hơn trong cùng một văn bản.

Tổng quát, phương pháp của Yarowsky (1995) [9] được tóm tắt như sau :

- B1 - Gọi F_k là tập các cụm từ đồng hiện thường xuất hiện với sắc thái ý nghĩa s_k. Ban đầu, F_k được khởi tạo dựa trên các định nghĩa của *w* trong từ điển hay đơn giản là được nhập vào bằng tay.
- B2 - Gọi E_k là tập tất cả các ngữ cảnh *c* có thể ấn định sắc thái ý nghĩa s_k lên từ đa nghĩa *w*. Ban đầu E_k được khởi tạo rỗng.
- B3 - Với mọi sắc thái ý nghĩa s_k ta xây dựng $E_k = \{c \mid \exists f : f \in c \wedge f \in F_k\}$
- B4 - Với mọi sắc thái ý nghĩa s_k ta xây dựng $F_k = \{ f \mid \forall s' \neq s_k : \frac{P(s_k \mid f)}{P(s' \mid f)} > \alpha \}$

  Lưu ý : α là ngưỡng định trước.
- B5 - Quay lại bước (B3) chừng nào còn ít nhất một E_k thay đổi.
- B6 - Từ các tập E_k đã xác định, ta có thể chỉ ra ngay sắc thái ý nghĩa nào của *w* là chiếm đa số trên toàn văn bản và đó chính là sắc thái ý nghĩa sẽ được gán cho từ đa nghĩa *w*.

Thực nghiệm cho thấy cách làm này tuy đơn giản nhưng đạt hiệu quả rất đáng ngạc nhiên. Những phiên bản khác nhau của mô hình này cho độ chính xác khoảng từ 90.5% đến 96.5%. Kết quả như vậy là rất khả quan vì mô hình hoàn toàn chỉ dùng từ điển và ngữ liệu đơn ngữ chứ không hề dùng ngữ liệu có gán nhãn ngữ nghĩa.

3.1.2 KHỬ NHẬP NHẰNG NGỮ NGHĨA CÓ GIÁM SÁT

Trong một mô hình học khử nhập nhằng ngữ nghĩa có giám sát, ngữ liệu tinh chế (có gán nhãn ngữ nghĩa) là bắt buộc phải có. Với mỗi từ đa nghĩa *w* một tập các

mẫu học sẽ được rút trích ra từ ngữ liệu bằng cách liệt kê tất cả các lần xuất hiện của *w* cùng nhãn ngữ nghĩa và ngữ cảnh tương ứng trên toàn bộ ngữ liệu. Khi đó, bài toán khử nhập nhằng ngữ nghĩa có thể xem như bài toán phân loại với vector đặc trưng là ngữ cảnh, các phân lớp ứng với các lớp ngữ nghĩa. Như vậy, mục đích cuối cùng của chúng ta là xây dựng một bộ phân loại có khả năng nhận vào ngữ cảnh của một từ đa nghĩa dưới dạng vector đặc trưng rồi chỉ ra sắc thái ý nghĩa thích hợp với nó nhất.

Ở đây chúng ta sẽ khảo sát hai mô hình học giám sát kinh điển : Bayesian Classification (Gale et al., 1992 [5]) và Information Theory (Brown et al., 1991 [2]). Hai mô hình này sử dụng các nguồn thông tin khác nhau trích từ ngữ cảnh và đều sử dụng khá hiệu quả. Mô hình Bayesian Classification xem ngữ cảnh như một dạng túi từ (bag of words), nghĩa là không quan tâm đến cấu trúc ngữ pháp giữa các từ mà chỉ tích hợp thông tin và quan hệ ngữ nghĩa giữa các từ trong cùng cửa sổ ngữ cảnh. Mô hình Information Theory thì chỉ tập trung phân tích các đặc trưng chứa thông tin hữu ích trong cửa sổ ngữ cảnh, theo ý nghĩa là chỉ những từ được xem là chứa đựng những thông tin quan trọng mới được lấy vào vector đặc trưng.

3.1.2.1 MÔ HÌNH PHÂN LOẠI BAYES

Ý tưởng của mô hình này là nhằm tính các xác suất $P_w(s|c)$ (xác suất để khi từ đa nghĩa *w* được đặt trong ngữ cảnh *c* thì nó mang ý nghĩa *s*). Nguyên tắc phân lớp ở đây là tối thiểu hóa sai số (Bayesian decision rule). Cụ thể, với một từ đa nghĩa *w* và một ngữ cảnh *c* cho trước, bộ phân loại sẽ gán sắc thái ý nghĩa *s'* cho từ *w* theo tiêu chí sau đây :

$$s' = argmax_s P_w(s|c) \quad (4.1)$$

Từ đó, vấn đề còn lại là làm sao tính được $P_w(s|c)$. Theo công thức Bayes ta có :

$$P_w(s|c) = P_w(s)P_w(c|s)/P_w(c) \quad (4.2)$$

Do $P_w(c)$ không đổi với mọi *s* nên ta có thể bỏ qua khi đó :

$$P_w(s|c) = P_w(s)P_w(c|s) \quad (4.3)$$

Giả thiết $c = v_1 v_2 ... v_N$, ta có thể giả sử v_i là độc lập có điều kiện với nhau (conditionally independent – *Naïve Bayes assumption*). Khi đó, ta có thể viết lại (4.3) như sau :

$$P_w(s|c) = P_w(s)P_w(c|s) = P_w(s)\prod_{i=1}^{N} P(v_i \mid s) \quad (4.4)$$

Để tiện xử lý khi lập trình, thay vì làm việc với $P_w(s|c)$ ta làm việc với $lnP_w(s|c)$. Như vậy, (4.1) sẽ được viết gọn lại như sau :

$$s' = argmax_s \left[lnP_w(s) + \sum_{i=1}^{N} \ln P_w(v_i \mid s) \right] \quad (4.5)$$

Trong đó :

- $P_w(s) = C_w(s) / C(w)$
- $P_w(v_i|s) = C_w(v_i|s) / C_w(s)$
- $C_w(s)$ là số lần từ đa nghĩa *w* xuất hiện với sắc thái ý nghĩa *s* trong ngữ liệu.
- $C_w(v|s)$ là số lần *v* xuất hiện trong ngữ cảnh của từ *w* với sắc thái ý nghĩa *s*.
- $C(w)$ là số lần từ đa nghĩa *w* xuất hiện trong ngữ liệu.

Ưu điểm của phương pháp này là nó đơn giản, dễ cài đặt và thực thi nhanh. Tuy nhiên, khuyết điểm của phương pháp này là nó đã dựa trên hai giả thiết không phải lúc nào cũng đúng trong ngôn ngữ :

- Thứ tự tuyến tính và cấu trúc giữa các từ trong ngữ cảnh có thể bỏ qua.
- Các từ xuất hiện trong cùng cửa sổ ngữ cảnh độc lập xác suất với nhau.

Theo Gale, Church, Yarowsky (1992) [5], phương pháp này cho độ chính xác khoảng 90% khi kiểm định trên tập các danh từ đa nghĩa : *duty, drug, land, language, position, sentence.*

3.1.2.2 MÔ HÌNH INFORMATION THEORY

Cơ sở của phương pháp này dựa trên khái niệm về độ đo thông tin tương hỗ :

$$I(X;Y) = \sum_{x \in X} \sum_{y \in Y} p(x,y) \ln \frac{p(x,y)}{p(x)p(y)} \quad (5.1)$$

Nhắc lại là ý nghĩa của độ đo này là thể hiện khả năng biết chính xác về giá trị của một biến ngẫu nhiên khi biết giá trị của biến ngẫu nhiên còn lại. Quay lại vấn đề của chúng ta giả thiết một từ *w* trong ngôn ngữ nguồn có hai sắc thái ý nghĩa khác nhau được dùng tùy thuộc ngữ cảnh (giả sử từ *w* chỉ có hai sắc thái ý nghĩa chỉ để tiện cho việc trình bày, trên thực tế thuật toán vẫn có thể mở rộng dễ dàng trong trường hợp *w* có nhiều hơn hai sắc thái ý nghĩa) và khi dịch sang ngôn ngữ đích *w* có thể được dịch bằng một trong các từ thuộc tập $T = \{t_1, t_2, ..., t_N\}$. Ý tưởng áp dụng trong phương pháp này là trong số các từ nằm trong cửa sổ ngữ cảnh của *w* có một số ít sẽ chứa những thông tin đáng tin cậy để quy định sắc thái ý nghĩa tương ứng của *w*.

Ví dụ :

- I plant a tree in my garden. (1)
- This plant is fading. (2)

Giả thiết cửa sổ ngữ cảnh là các từ xung quanh từ đa nghĩa *plant* trong phạm vi [-2; +2] thì ngữ cảnh trong câu (1) là {*I, a, tree*} còn ngữ cảnh trong câu (2) là {*This, is, fading*}. Ta nhận xét trong cả hai ngữ cảnh từ đứng trước (v_{-1}) *plant* chứa thông tin tuyệt đối để xác định ngữ nghĩa của nó. Nếu đó là chủ từ (*he, she, I, we, they*) thì *plant* phải hiểu là động từ mang ý nghĩa là một hành động trồng trọt còn nếu đó là giới từ (*a, the*) hay tính từ chỉ định (*this, that*) thì *plant* phải hiểu là danh từ chỉ thực vật. Như thế, rõ ràng từ đứng trước (v_{-1}) từ đa nghĩa *plant* chứa lượng thông tin đáng tin cậy để khử nhập nhằng ngữ nghĩa hơn hẳn những từ khác trong ngữ cảnh.

Vấn đề là làm sao để biết từ nào có tiềm năng khử nhập nhằng ngữ nghĩa hơn những từ còn lại ? Để tiện cho việc trình bày ta tạm gọi những từ như thế là "từ chỉ định" (indicator). Như vậy, để biết được từ nào là từ tiềm năng thì ta cần phải ước tính "tiềm năng" của từng từ. Để ước tính tiềm năng đó thì ta có thể sử dụng độ đo thông tin tương hỗ đã nói ở trên với ý nghĩa là khả năng nhận biết chính xác sắc thái ý nghĩa khi biết thông tin về từ chỉ định đó. Do đó, độ đo này càng lớn thì tiềm năng của từ chỉ định càng cao. Việc tính độ đo tiềm năng đó được thực hiện như sau :

- B0 - Gọi $X = \{ x_1, x_2, ..., x_M \}$ là tập các từ trong ngôn ngữ đích mà từ chỉ định v_i có thể được dịch sang.
- B1 - Đầu tiên ta phân hoạch T một cách ngẫu nhiên thành $\{T_1, T_2\}$.
- B2 - Tiếp theo, ta phân hoạch X thành $\{X_1, X_2\}$ (ứng với T mới phân hoạch) sao cho $I(T; X)$ cực đại.
- B3 - Phân hoạch lại T sao cho $I(T; X)$ cực đại (ứng với X mới phân hoạch lại).
- B4 - Lặp lại bước B2 chừng nào độ tăng của $I(T; X)$ còn lớn hơn một ngưỡng cụ thể.

Lưu ý : khi ta phân hoạch một tập $R = \{r_1, r_2, ..., r_N\}$ thành $R = \{R_1, R_2\}$ thì khi đó xem như tập R chỉ có hai giá trị s_1, s_2 ứng với R_1 và R_2. Khi đó, $p(s_1) = |R_1|/|R|$ và $p(s_2) = |R_2|/|R|$.

Với mỗi từ trong cửa sổ ngữ cảnh ta tính tiềm năng cho nó theo cách trên và chọn từ có tiềm năng cao nhất làm từ chỉ định. Khi đó, xét trong cửa sổ ngữ cảnh nếu từ chỉ định v_i mang giá trị $x_i \in X_1$ thì w sẽ mang sắc thái ý nghĩa thứ nhất và ngược lại. Thực nghiệm cho thấy phương pháp trên khá hiệu quả và dễ dàng tích hợp vào một hệ dịch máy tự động do bản thân mô hình khử nhập nhằng ngữ nghĩa này sử dụng ngữ liệu song ngữ có gán nhãn ngữ nghĩa trên ngôn ngữ nguồn. Theo Brown et al. (1984) phương pháp này cải thiện chất lượng dịch máy lên 20% (từ 37 câu đúng lên

45 câu đúng trên tổng số 100 câu). Tuy nhiên, nhược điểm của phương pháp này là nó đòi hỏi quá nhiều về phương diện ngữ liệu nên khó có thể triển khai trên diện rộng.

3.1.3 KHỬ NHẬP NHẰNG NGỮ NGHĨA KHÔNG GIÁM SÁT

Trong nhiều trường hợp, ngữ liệu tinh chế để học giám sát rất ít hay không có sẵn, chẳng hạn như khi ta tiến hành khử nhập nhằng ngữ nghĩa trên những văn bản chuyên ngành. Lúc đó, thậm chí ta còn có thể không biết rõ một từ chuyên ngành có thể có bao nhiêu sắc thái ý nghĩa. Vì thế, tài nguyên mà chúng ta có chỉ là những kho ngữ liệu thô như văn bản song ngữ, đơn ngữ, từ điển định nghĩa các sắc thái ý nghĩa, … Trong những tình huống như vậy, chúng ta có thể tiếp cận theo phương pháp học không giám sát trực tiếp (dựa trên ngữ liệu đơn ngữ) hay là gián tiếp (dựa trên ngữ liệu song ngữ).

Ý tưởng của phương pháp học không giám sát thuần túy là kết hợp giải thuật ước lượng trung bình cực đại EM (Expectation – Maximization) để xác định tham số cho một mô hình học mẫu thông thường nào đó. Ý tưởng của phương pháp học không giám sát dựa trên ngữ liệu song ngữ thì dựa trên nhận xét là những ngữ nghĩa khác nhau của từ đa nghĩa trong ngôn ngữ nguồn thường sẽ thể hiện khác nhau trên ngôn ngữ đích. Khi đó, mục tiêu của phương pháp học này chính là tìm cách ước lượng liên kết từ giữa ngôn ngữ nguồn và ngôn ngữ đích dựa trên ngữ liệu song ngữ. Liên kết từ đó sẽ hỗ trợ chúng ta xây dựng ngữ liệu có gán nhãn ngữ nghĩa (một cách tương đối) từ ngữ liệu song ngữ và sau đó, chúng ta có thể tiến hành học có giám sát theo cách thông thường. Trong phần còn lại của chương này, chúng ta sẽ khảo sát chi tiết từng phương pháp nói trên.

3.1.3.1 PHƯƠNG PHÁP TRỰC TIẾP

Về nguyên tắc, khi ta phải kết hợp giải thuật ước lượng cực đại cho các tham số của một mô hình nào đó nghĩa là chúng ta không có đủ thông tin để ước lượng các tham số đó một cách trực tiếp và vì thế, chúng ta sẽ tìm cách ước lượng giá trị trung bình cực đại cho các tham số đó. Theo cách làm này, việc đầu tiên ta làm tạo ngẫu nhiên một bộ giá trị cho các tham số sau đó từ từ cải thiện chúng để cực đại hóa hàm log likelihood. Hàm log likelihood, như chúng ta sẽ thấy trong minh họa dưới đây, chính là phản ánh độ "khớp" của mô hình đặc trưng bởi bộ giá trị hiện hành của các tham số với ngữ liệu thô. Như vậy việc cải thiện hàm log likelihood theo hướng tăng dần cũng chính là cải thiện mô hình học "khớp" dần với ngữ liệu.

Để minh họa, ta sẽ dùng lại mô hình Bayesian Classification (Gale et al. 1992 [5]) nói trên kết hợp với giải thuật EM :

- B1 - Khởi tạo bộ tham số ngẫu nhiên $P_w(v_j|s_k)$ và $P_w(s_k)$
- B2 - Tính hàm log likelihood của ngữ liệu ứng với mô hình vừa khởi tạo ngẫu nhiên $l(C) = \ln \prod \sum P_w(c_i \mid s_k) P_w(s_k) = \sum \log \sum P_w(c_i \mid s_k) P_w(s_k)$ (5.2)
- B3 - Ước lượng $h_{i,k}$ là xác suất hậu nghiệm (posterior probability) để sắc thái ý nghĩa s_k của từ w được thể hiện qua ngữ cảnh c_i : $h_{i,k} = \frac{P_w(c_i \mid s_k)}{\sum P_w(c_i \mid s_k)}$ (5.3)

  Lưu ý : để tính $P_w(c_i|s_k)$ thì ta có thể sử dụng giả thiết Naive Bayes giống như trong mô hình học có giám sát nói trên.
- B4 - Từ $h_{i,k}$ ta ước lượng lại bộ tham số $P_w(v_j|s_k)$, $P_w(s_k)$ nhằm cải thiện hàm log likelihood :

$$P_w(v_j \mid s_k) = \frac{\sum_{\{c_i : v_j \in c_i\}} h_{i,k}}{\sum_k \sum_{\{c_i : v_j \in c_i\}} h_{i,k}} \quad (5.4)$$

$$P_w(s_k) = \frac{\sum_i h_{i,k}}{\sum_k \sum_i h_{i,k}} \quad (5.5)$$

- B5 - Lặp lại bước B3 chừng nào độ tăng của hàm log likelihood còn lớn hơn ngưỡng.

Một vấn đề cần phải xem xét từ hướng tiếp cận này là khi ta không biết chính xác số lượng sắc thái ý nghĩa của từ *w* thì ta sẽ ước lượng tham số *K* ra sao. Phương pháp phổ biến thường dùng trong trường hợp này là thăm dò. Cụ thể là ta sẽ bắt đầu với một giá trị *K* đủ nhỏ sau đó ta từ từ tăng *K* lên dần dần. Tại một thời điểm, nếu việc tăng giá trị *K* lên làm hàm log likelihood tăng lên đáng kể thì ta tiếp tục tăng K, ngược lại thì ta kết thúc tại giá trị *K* hiện hành.

3.1.3.2 PHƯƠNG PHÁP GIÁN TIẾP

Đối với phương pháp học không giám sát gián tiếp, chúng ta dựa trên cơ sở là ngữ liệu song ngữ để xây dựng ngữ liệu có gán nhãn thông qua các mô hình liên kết từ (xem Chương 2). Khi đã xây dựng được ngữ liệu có gán nhãn ngữ nghĩa thì ta đưa về bài toán phân loại thông thường và áp dụng các mô hình học giám sát để giải quyết. Để minh họa cho phương pháp này, sau đây chúng ta sẽ tìm hiểu mô hình xử lý nhập nhằng ngữ nghĩa dựa trên ngữ liệu song ngữ theo nghiên cứu của Vickrey (2005) [25].

Trước hết, quan điểm của nghiên cứu này cho rằng bài toán xử lý nhập nhằng ngữ nghĩa thực chất là bài toán trung chuyển thường được phát sinh trong ngữ cảnh một bài toán lớn nào đấy, chẳng hạn như dịch máy. Theo đó, việc xem xét bài toán xử lý nhập nhằng như là một bài toán độc lập sẽ tự tạo ra những vấn đề phức tạp mà chúng ta khó có thể giải quyết cho ổn thỏa được. Trên thực tế, hiện nay có rất nhiều công trình nghiên cứu về vấn đề xử lý nhập nhằng ngữ nghĩa nhưng hầu hết chỉ xem

xét nó như một bài toán độc lập mà chưa quan tâm đến việc sẽ tích hợp nó vào ngữ cảnh của các bài toán lớn như thế nào.

Để dễ hình dung, chúng ta có thể lấy ví dụ từ bài toán dịch máy. Đối với bài toán dịch máy, việc xử lý nhập nhằng ngữ nghĩa là thiết yếu nhưng nếu ta tiếp cận theo cách truyền thống là phân loại ngữ nghĩa của từ đa nghĩa dựa trên từ điển ngữ nghĩa cho trước thì sẽ rất khó khăn khi tích hợp vào mô hình dịch máy. Vấn đề ở đây là nếu đặt trong ngữ cảnh của bài toán lớn là dịch máy thì bản thân hệ thống nhãn ngữ nghĩa phải được, bằng cách nào đấy, ánh xạ lên ngôn ngữ đích và việc ánh xạ như thế sẽ đòi hỏi rất nhiều công sức và thời gian.

Như vậy, để khắc phục hạn chế đó, giải pháp được đề xuất là sử dụng tập hợp các văn bản song ngữ. Cách tiếp cận này có lợi điểm là tài nguyên song ngữ vào thời điểm hiện tại khá phổ biến nên ta không sợ thiếu và ngoài ra, việc xây dựng hệ thống nhãn ngữ nghĩa trên cơ sở của liên kết từ sẽ giúp ta giải quyết một cách tự nhiên vấn đề ánh xạ nhãn ngữ nghĩa lên ngôn ngữ đích. Hơn thế nữa, cách tiếp cận dựa trên ngữ liệu song ngữ sẽ có khả năng thích nghi tốt đối với tính không ổn định của ngôn ngữ. Trong phần còn lại của mục này, chúng ta sẽ mô tả chi tiết phương pháp này qua các phần xây dựng ngữ liệu có gán nhãn ngữ nghĩa và huấn luyện mô hình học.

Đầu tiên, để thực hiện liên kết từ trên ngữ liệu song ngữ, phương pháp này sử dụng công cụ GIZA++ (xem Phụ lục C.3) được cài đặt dựa trên cơ sở lý thuyết các mô hình liên kết từ đã trình bày ở chương trước. Cụ thể, chúng ta gọi *S* ngôn ngữ nguồn và *T* là ngôn ngữ đích. Khi tiến hành liên kết từ theo hướng $S \to T$, GIZA++ sẽ liên kết mỗi từ trong *T* với tối đa một từ trong *S*. Khi đó, với mỗi từ *a* trong *S*, chúng ta có tập $U_{a,S\to T} = \{b_1, b_2, ..., b_k\}$ mà trong đó b_i là một từ hay cụm từ trong *T* được liên kết với *a*. Ngoài ra, để hạn chế nhiễu trong $U_{a,S\to T}$ chúng ta chạy liên kết từ theo hướng $T \to S$ (mỗi từ *a* trong *S* sẽ liên kết với tối đa một từ trong *T*). Lúc đó, với

mỗi $b_i \in U_{a,S\to T}$, ta ràng buộc rằng mọi từ $w \in b_i$ thì $U_{w,T\to S} = \{a\}$ hay $|U_{w,T\to S}| = 0$. Ngoài ra, để nâng cao tính chính xác, ta còn ràng buộc số lần xuất hiện của mỗi b_i phải lớn hơn một ngưỡng tối thiểu định trước. Cuối cùng, ta thu được tập U_a chỉ chứa những b_i thỏa các điều kiện nói trên được chọn làm tập các nhãn ngữ nghĩa trên ngôn ngữ đích cho từ đa nghĩa *a*.

Ta gọi $D_{a,S}$ là tập các câu trong ngôn ngữ nguồn *S* có chứa từ đa nghĩa *a* và định nghĩa tập dữ liệu huấn luyện khử nhập nhằng ngữ nghĩa cho *a* là $C_a = \{s \in D_{a,S}, tag = U_{a,s\to t}\}$. Từ đó, với mỗi từ *a*, chúng ta sẽ có một mô hình khử nhập nhằng ngữ nghĩa riêng được huấn luyện dựa trên C_a. Về nguyên tắc, khi đã xây dựng được C_a, ta có thể áp dụng bất kỳ mô hình học phân loại nào như Bayes Classification, Information Theory, … Ở đây, mô hình được sử dụng trong nghiên cứu của Vickrey (2005) là hồi quy tuyến tính :

$$P_{\theta_a}(U_{a,s} = b \mid s, a) = \frac{e^{\theta_a^b \varphi^{a,s}}}{\sum_{b' \in U_a} e^{\theta_a^{b'} \varphi^{a,s}}} \quad (6.1)$$

Trong đó, ta có :

- $\theta_a = \{\theta_a^b \mid b \in U_a\}$ là tập các tham số của mô hình hồi quy tuyến tính, các tham số này sẽ được ước lượng dựa trên tập huấn luyện C_a.
- $\varphi^{a,s}$ là vector đặc trưng rút trích từ ngữ cảnh trong câu *s* xung quanh từ đa nghĩa *a* (bao gồm thông tin về POS(*a*), những từ nằm trong cửa sổ ngữ cảnh).
- θ_a^b là tham số vector ứng với nhãn ngữ nghĩa *b* của từ đa nghĩa *a* và như thế, ta có $|\theta_a^b| = |\varphi^{a,s}|$.

Để ước lượng bộ các tham số θ_a cho mô hình hồi quy tuyến tính nói trên, cách phổ biến là cực đại hóa hàm log likelihood ứng với ngữ liệu huấn luyện :

$$L(D_a) = \sum_{\{s,tag\} \in C_a} \ln P_{\theta_a}(tag \mid a, s) \quad (6.2)$$

Để cực đại hóa hàm log likelihood nói trên, ta có thể áp dụng phương pháp Conjugate Gradient Ascent (Shewchuk, 1994).

3.2 MÔ HÌNH ĐỀ XUẤT

Trong luận văn này, mô hình khử nhập nhằng ngữ nghĩa được xây dựng dựa trên ý tưởng của những hướng tiếp cận trước đó. Về mặt cấu trúc, mô hình này sẽ được chia làm ba phần riêng biệt : xây dựng ngữ liệu có gán nhãn ngữ nghĩa từ văn bản song ngữ, huấn luyện mô hình học mẫu có giám sát, khai thác ngữ liệu đơn ngữ để mở rộng ngữ liệu huấn luyện và nâng cấp chất lượng cho mô hình phân loại. Trong các tiểu mục tiếp theo của chương này, chúng ta sẽ lần lượt mô tả chi tiết từng phần nói trên.

3.2.1 XÂY DỰNG NGỮ LIỆU TỪ VĂN BẢN SONG NGỮ

Lấy ý tưởng của các phương pháp tiếp cận trước, mô hình này cũng sử dụng ngữ liệu song ngữ để xây dựng dữ liệu huấn luyện thông qua liên kết từ. Như đã đề cập đến ở các chương trước, liên kết từ trong ngữ liệu song ngữ có hai phương pháp phổ biến là liên kết từ theo lớp ngữ nghĩa và liên kết từ thống kê. Cả hai phương pháp đều có những ưu khuyết điểm riêng : liên kết từ thống kê thì cho độ phủ cao nhưng độ chính xác thấp trong khi liên kết từ theo lớp ngữ nghĩa thì ngược lại.

Tuy nhiên, chúng ta có nhận xét là đối với vấn đề thu thập dữ liệu cho việc xử lý nhập nhằng ngữ nghĩa với một từ đa nghĩa *w* nào đó thì chúng ta chỉ cần liên kết từ cho chính bản thân *w* chứ không cần quan tâm đến những từ khác. Do đó, khi liên kết từ thay vì tiến hành cùng lúc với mọi từ như các mô hình liên kết từ nói trên,

chúng ta chỉ cần tập trung vào tìm kiếm liên kết từ tương ứng cho *w*. Với cách tiếp cận như vậy, độ phủ của liên kết từ sẽ được cải thiện đáng kể do chúng ta không phải quan tâm đến liên kết từ cho những từ dễ gây nhiễu như trợ từ, hư từ, từ chức năng mà chỉ tập trung tìm kiếm liên kết từ cho *w*.

Ở trong luận văn này, phương pháp liên kết từ cho từ đa nghĩa *w* sẽ được xây dựng dựa trên ý tưởng của thuật toán *DictAlign* trong mô hình liên kết từ theo lớp ngữ nghĩa. Lý do chúng ta xây dựng phương pháp liên kết từ cho từ đa nghĩa *w* dựa trên mô hình liên kết từ theo lớp ngữ nghĩa là vì khi độ phủ đã đảm bảo (như đã phân tích ở trên) thì chúng ta cần độ chính xác cao mà để đạt được độ chính xác cao thì không thể chỉ dựa vào thống kê trên ngữ liệu song ngữ.

Về phương diện tài nguyên, ngoài hệ thống nhãn ngữ nghĩa LLOCE cho tiếng Anh, chúng ta còn có thêm tập hợp các lớp từ đồng nghĩa tiếng Việt, tạm gọi là LLOCV, là dạng từ điển đồng nghĩa được xây dựng dựa trên LLOCE. Cụ thể, ứng với mỗi lớp từ đồng nghĩa trong LLOCE thì chúng ta cũng có một lớp từ đồng nghĩa tương ứng bên LLOCV (hai từ điển LLOCE và LLOCV sử dụng cùng một cấu trúc nhãn ngữ nghĩa). Trước khi đi vào chi tiết giải thuật, chúng ta sẽ giới thiệu và thống nhất trên một số ký pháp và khái niệm liên quan :

- Định nghĩa $C_w = \{ (s, t) \mid s \supset w \}$ là ngữ liệu song ngữ dùng để huấn luyện mô hình khử nhập nhằng ngữ nghĩa cho từ đa nghĩa *w* với *s* là câu nguồn (tiếng Anh) chứa *w*, *t* là bản dịch của *s* trên ngôn ngữ đích (tiếng Việt).

- Định nghĩa L_w là tập các nhãn ngữ nghĩa mà *w* có thể có (do một từ đa nghĩa có thể thuộc nhiều nhóm ngữ nghĩa).

- Định nghĩa V_x là tập các từ đồng nghĩa tiếng Việt trong lớp ngữ nghĩa với nhãn $x \in$ LLOCV.

- Định nghĩa *Sim(dw, dt)* là độ tương tự của hai từ *dw* và *dt* trong tiếng Việt :

$$Sim(dw, dt) = \frac{2 \mid dw \cap dt \mid}{\mid dw \mid + \mid dt \mid}$$

  Trong đó, |*dw*|, |*dt*|, $\mid dw \cap dt \mid$ lần lượt là số âm tiết tiếng Việt xuất hiện trong *dw*, *dt* và cả hai từ *dw* và *dt*.

- Định nghĩa $score_w(v)$ là điểm ứng viên làm từ liên kết tương ứng của *w* đối với từ tiếng Việt *v*.

Giải thuật xây dựng ngữ liệu có gán nhãn ngữ nghĩa cho từ đa nghĩa *w* từ tập văn bản song ngữ :

- B1 : Khởi tạo $D_w = \bigcup_{x \in L_w} V_x$, $score_w(v) = 0 \forall v$, P_w = {}. Trong đó, D_w là tập tất cả các từ thuộc các lớp đồng nghĩa với *w* trong LLOCV.

- B2 : Với mỗi cặp câu *p* = (*s*, *t*) trong ngữ liệu song ngữ Anh – Việt C_w, ta lần lượt thực hiện các bước sau :

  - B2.1 : Tiến hành tách từ trên *t* và loại bỏ đi những từ có tần suất xuất hiện cao (trợ từ, hư từ, từ chức năng) trong tiếng Việt để thu được danh sách từ L_t.

  - B2.2 : Khởi tạo danh sách ứng viên $Q_{p,w}$ = {}.

  - B2.3 : Với mỗi từ $v \in L_t$, ta tính $Sim(v, D_w) = \max_{r \in D_w} Sim(v, r)$. Nếu $Sim(v, D_w) \geq h_1$ (ngưỡng tối thiểu cho độ tương tự), ta cập nhật danh sách ứng viên $Q_{p,w} = Q_{p,w} + \{v\}$.

- B2.4 : Với mỗi từ $v \in Q_{p,w}$, ta cập nhật điểm ứng viên cho v : $score_w(v) = score_w(v) + 1$.

- B3 : Với mỗi cặp câu p = (s, t) trong ngữ liệu song ngữ Anh – Việt C_w, ta lần lượt thực hiện các bước sau :

  - B3.1 : Điều chỉnh $Q_{p,w} = \{v \in Q_{p,w} \mid score_w(v) \geq h_2\}$, nếu $Q_{p,w}$ = {} thì ta bỏ qua p.

  - B3.2 : Xác định liên kết từ thích hợp nhất với w trong ngữ cảnh cặp câu Anh – Việt $v' = \arg\max_{v \in Q_{p,w}} \{score_w(v) \mid Sim(v, D_w) = \max_{v \in Q_{p,w}} Sim(v, D_w)\}$.

  - B3.3 : Xác định nhãn ngữ nghĩa $l = \arg\max_{l' \in L_w} Sim(v', V_{l'})$.

  - B3.4 : Cập nhật kết quả : $P_w = P_w + \{(s, l)\}$.

- B4 : Output kết quả P_w = {(s, l) | $l \in L_w$} là tập các mẫu học để khử nhập nhằng ngữ nghĩa cho từ w, mỗi mẫu học bao gồm ngữ cảnh tương ứng s và nhãn ngữ nghĩa l.

3.2.2 XÂY DỰNG MÔ HÌNH PHÂN LOẠI NGỮ NGHĨA

Một trong những phương pháp phổ biến thường được áp dụng cho bài toán khử nhập nhằng ngữ nghĩa xây là mô hình học phân loại Bayes (Naïve Bayesian Classification) đề xuất bởi Gale et al. (1992) [5]. Đây là phương pháp phân loại điển hình ứng dụng lý thuyết thống kê trong xử lý. Nguyên tắc khử nhập nhằng ngữ nghĩa của phương pháp này là chọn ngữ nghĩa s của từ đa nghĩa w sao cho xác suất

$P_w(s \mid F)$ đạt giá trị lớn nhất. Trong đó, $F = \{f_1, f_2, ..., f_N\}$ là vector đặc trưng rút trích từ ngữ cảnh của *w* với giả thiết là các f_i độc lập xác suất với nhau. Khi đó, ta có thể viết :

$$s' = \arg\max_s \ln P_w(s \mid F) = \arg\max_s \ln\{P_w(F \mid s)P_w(s)\} \quad (7.1)$$

$$P(F \mid s) = \prod_{i=1}^{N} P(f_i \mid s) \quad (7.2)$$

Từ công thức (7.2), ta có thể viết lại (7.1) như sau :

$$s' = \arg\max_s \ln P_w(F \mid s)P_w(s) = \arg\max_s \{\sum_{i=1}^{N} \ln P_w(f_i \mid s) + \ln P_w(s)\} \quad (7.3)$$

Các giá trị xác suất $P_w(f \mid s)$ và $P_w(s)$ là các tham số của mô hình phân loại Bayes và được xác định theo nguyên tắc ước lượng cực đại (MLE) từ ngữ liệu huấn luyện:

- $P_w(s) = \frac{|C_w(s)|}{|C_w|}$ với C_w là tập huấn luyện của từ *w* và $C_w(s)$ là tập hợp các mẫu học với nhãn ngữ nghĩa *s* trong C_w.

- $P_w(f \mid s) = \frac{|C_w(f,s)|}{|C_w(s)|}$ với $C_w(f,s)$ là tập hợp các mẫu học có sự xuất hiện của *f* trong ngữ cảnh và được gán nhãn ngữ nghĩa *s*.

Lưu ý : Trong quá trình xử lý nhập nhằng ngữ nghĩa, nếu mô hình gặp phải đặc trưng *f* chưa từng xuất hiện trong tập huấn luyện thì có thể tạm gán $P_w(f \mid s) = \frac{1}{|C_w|}$ để làm trơn các giá trị tham số.

Mô hình Naïve Bayesian Classification nói trên đã được thử nghiệm và so sánh với các mô hình học giám sát khác trên cùng bài toán khử nhập nhằng ngữ nghĩa (với

cùng bộ rút trích đặc trưng ngữ cảnh) trong nhiều công trình nghiên cứu. Kết quả thực nghiệm cho thấy mặc dù có kiến trúc đơn giản nhưng chất lượng của phương pháp phân loại Bayes hoàn toàn không thua kém những mô hình phức tạp khác. Điển hình của các nghiên cứu đó là báo cáo của Mooney (1996) [11] và Pedersen (2000) [14].

Mooney (1996) [11] so sánh hiệu quả của 6 mô hình học giám sát trong vấn đề khử nhập nhằng ngữ nghĩa bao gồm : Naïve Bayesian Classification, Perceptron, Decision Tree, K - Nearest Neighbour (KNN), Logic-based Disjunctive Normal Form (DNF), Conjunctive Normal Form (CNF) và kết luận rằng Naïve Bayesian Classification là phương pháp cho hiệu quả tốt nhất (tất cả các phương pháp nói trên đều sử dụng cùng vector đặc trưng rút trích từ tập các từ trong ngữ cảnh đang xét).

Pedersen (2000) [14] đề xuất phương pháp kết hợp nhiều mô hình Naïve Bayesian Classification và thu được kết quả rất khả quan. Qua đó, Pedersen chứng minh rằng để xây dựng một mô hình xử lý nhập nhằng ngữ nghĩa hiệu quả chúng ta có thể xuất phát từ cách kết hợp nhiều mô hình đơn giản như Naïve Bayesian Classification. Cụ thể, trong cách tiếp cận của mình, Pedersen đã kết hợp 9 mô hình phân loại Bayes với các cửa sổ ngữ cảnh kích thước 0, 1, 2, 3, 4, 5, 10, 20 và 50. Kết quả thử nghiệm trên hai tập dữ liệu của hai từ đa nghĩa "*line*" và "*interest*" là 89% và 88%.

Như vậy, trong luận văn này, mô hình dùng để xử lý nhập nhằng ngữ nghĩa được chọn là mô hình Naïve Bayesian Classification. Tuy nhiên, khác với những mô hình phân loại Bayes thông thường vốn chỉ chú trọng đến việc khai thác thông tin từ sự xuất hiện của một số từ trong cửa sổ ngữ cảnh, mô hình này tận dụng thông tin từ nhiều khía cạnh để nâng cao độ chính xác.

Về nguyên tắc, để xử lý nhập nhằng ngữ nghĩa chúng ta chủ yếu dựa vào hai cơ sở chủ yếu là chủ đề của ngữ cảnh và mối liên hệ cấu trúc giữa từ cần khử nhập nhằng

ngữ nghĩa và các từ xung quanh. Để xác định chủ đề của ngữ cảnh, chúng ta có thể dựa vào tập hợp (không kể đến thứ tự) của các từ nằm trong cửa sổ ngữ cảnh lớn (toàn bộ câu đang xét hay vài câu xung quanh) và để phát hiện ra các mối liên hệ cấu trúc giữa những từ xung quanh với từ cần được khử nhập nhằng ngữ nghĩa chúng ta cần quan tâm đến các cụm từ đi liền nhau trong cửa sổ ngữ cảnh hẹp (vài từ xung quanh từ đang xét). Cụ thể, chúng ta sẽ xây dựng mô hình học phân loại dựa trên các tập đặc trưng rút từ ngữ cảnh như sau :

- F_1 = {…, w_{-2}, w_{-1}, w_1, w_2, …} là tập hợp (không kể đến thứ tự của các từ trong cửa sổ ngữ cảnh lớn – giả thiết từ cần khử nhập nhằng ngữ nghĩa là $w = w_0$).
- F_2 = {…, (w_{-2}, -2), (w_{-1}, -1), (w_1, 1), (w_2, 2), …} là tập hợp các từ cùng vị trí tương đối của chúng trong cửa sổ ngữ cảnh hẹp.
- F_3 = {$w_{-h}…w_{-1}w$, $w_{-(h-1)}…ww_1$, $w_{-(h-2)}…ww_1w_2$, …, $ww_1…w_h$} là tập hợp các cụm từ xung quanh w với độ dài h.
- F_4 = { $p_{-h}…p_{-1}w$, $p_{-(h-1)}…wp_1$, $p_{-(h-2)}…wp_1p_2$, …, $wp_1…p_h$} là tập hợp các cụm nhãn ngữ pháp xuất hiện xung quanh w trong ngữ cảnh hẹp độ rộng h.

Tóm lại, tập các đặc trưng ngữ cảnh mà mô hình phân loại Bayes đưa vào xử lý trong luận văn này là $F = F_1 \cup F_2 \cup F_3 \cup F_4$. Để dễ hình dung, chúng ta có thể xem xét ví dụ sau :

coil <NNS> up <IN> the <DT> dry <JJ> ***line*** *<NN> and <CC> stand <VB> midstream <NN>, <,> rod <NN> in <IN> instant <NN> readiness <NN> . <.>*

Giả thiết, ta lấy kích thước cửa sổ ngữ cảnh hẹp là 2 từ về phía trái, phải của từ đang xét (*line*). Lúc đó, ta có :

- F_1 = {*coil, dry, stand, midstream, rod, instant, readiness*}
- F_2 = {(*the, -2*), (*dry, -1*), (*and, 1*), (*stand, 2*)}

- F_3 = {*the dry line, dry line and, line and stand*}
- F_4 = {*<DT> <JJ> line, <JJ> line <CC>, line <CC> <VB>*}

Cuối cùng, $F = F_1 \cup F_2 \cup F_3 \cup F_4$ = {coil, dry, stand, midstream, rod, instant, readiness, (the, -2), (dry, -1), (and, 1), (stand, 2), the dry line, dry line and, line and stand, <DT> <JJ> line, <JJ> line <CC>, line <CC> <VB>}.

3.2.3 KHAI THÁC NGỮ LIỆU ĐƠN NGỮ

Sau khi được huấn luyện với ngữ liệu song ngữ như đã mô tả ở trên, mô hình khử nhập nhằng ngữ nghĩa sẽ được tăng cường bằng việc khai thác trên ngữ liệu đơn ngữ. Ý tưởng của việc tăng cường chất lượng cho mô hình khử nhập nhằng ngữ nghĩa xuất phát từ nhận xét của Yarowsky (1995) [9]:

- Nếu một từ đa nghĩa *w* xuất hiện nhiều lần trong cùng một văn bản hay một đoạn văn thì gần như tất cả các lần xuất hiện đó đều có cùng một sắc thái ý nghĩa. (One sense per discourse)

- Những từ hay cụm từ xuất hiện trong cửa sổ ngữ cảnh xung quanh từ đa nghĩa *w* cung cấp thông tin vững chắc và nhất quán về sắc thái ý nghĩa của *w*. (One sense per collocation)

Theo đó, ta có nhận xét là nếu một từ đa nghĩa *w* xuất hiện nhiều lần trong một văn bản cho trước và phần lớn số lần xuất hiện của *w* được mô hình phân loại ngữ nghĩa xây dựng ở trên gán cho ngữ nghĩa *s* thì nhiều khả năng là tất cả tất cả các lần xuất hiện của *w* đều mang nghĩa *s*. Với nhận xét đó, ta có thể sử dụng mô hình phân loại ngữ nghĩa đã được huấn luyện tốt với ngữ liệu song ngữ để tiến hành khai thác thêm các mẫu học trên ngữ liệu đơn ngữ - giả thiết rằng ngữ liệu đơn ngữ của chúng ta được tổ chức dưới dạng tập hợp các văn bản. Quy trình trên được mô tả chi tiết như sau :

- B1 : Xây dựng tập hợp S_w gồm tất cả các văn bản có chứa *w* từ kho ngữ liệu đơn ngữ.

- B2 : Chọn $D^* = \arg\max_{D \in S_w} score(D)$ - D^* là văn bản có độ ưu tiên cao nhất được xác định theo công thức sau đây :

$$score(D) = \sum_{f \in D} \delta(f, F_w) \text{ với } \delta(f, F_w) = \begin{cases} 1 : \sum_s P_w(f \mid s) \geq h \\ 0 : \sum_s P_w(f \mid s) < h \end{cases} \quad (7.4)$$

- B3 : Gọi *d(s)* là số lần nhãn ngữ nghĩa *s* được gán với các lần xuất hiện của *w* trong D^* bởi mô hình khử nhập nhằng ngữ nghĩa hiện tại, ta chọn $s^* = \arg\max_s d(s)$ làm nhãn ngữ nghĩa cho tất cả các lần xuất hiện của *w* trong D^*.

- B4 : Gọi E^* là tập các ngữ cảnh ứng với những lần xuất hiện của *w* trong D^*, ta xây dựng tập mẫu học bổ sung $E_w = \{(e, s^*) \mid e \in E^*\}$.

- B5 : Cập nhật các tham số $P_w(f \mid s)$ và $P_w(s)$ ứng với việc học bổ sung E_w - tương đương với việc huấn luyện lại mô hình phân loại ngữ nghĩa từ đầu với tập huấn luyện $E_w + C_w$

- B6 : Cập nhật S_w, C_w : $S_w = S_w - \{D^*\}, C_w = C_w + E_w$. Quay lại B2 chừng nào còn thỏa mãn điều kiện $|S_w| > 0$.

Như vậy, khi kết thúc giải thuật học tăng cường trên, ta thu được mô hình phân loại ngữ nghĩa được gia cố chất lượng trên ngữ liệu đơn ngữ và kho ngữ liệu có gán nhãn ngữ nghĩa đã được mở rộng. Kho ngữ liệu này có thể được tái sử dụng trong các nghiên cứu khác có liên quan đến vấn đề ngữ nghĩa của từ.

CHƯƠNG 4 – KẾT QUẢ THỬ NGHIỆM

4.1 XÂY DỰNG NGỮ LIỆU TỪ VĂN BẢN SONG NGỮ

Đối với phương pháp trình bày trong luận văn này, bộ ngữ liệu có gán nhãn ngữ nghĩa dùng làm dữ liệu huấn luyện cho mô hình phân loại ngữ nghĩa được xây dựng từ một phần của kho ngữ liệu song ngữ EVC (gồm 60,000 cặp câu song ngữ Anh – Việt) phát triển bởi nhóm VCL. Do ngữ liệu song ngữ sử đưa vào dụng còn tương đối hạn chế nên trong phạm vi của luận văn này, các thí nghiệm chỉ tiến hành trên phạm vi các danh từ đa nghĩa *bank, arm, plant, interest*. Ở đây, xin được nói rõ là luận văn này không hướng tới việc xây dựng một mô hình xử lý nhập nhằng ngữ nghĩa hoàn chỉnh mà chỉ nhằm chứng minh bằng thực nghiệm tính khả thi của một hướng đi liên quan đến vấn đề này.

Để chứng minh bằng thực nghiệm hiệu quả của phương pháp xây dựng ngữ liệu đề xuất trong luận văn này, chúng ta sẽ so sánh kết quả xây dựng ngữ liệu có gán nhãn của phương pháp này với phương pháp ứng dụng liên kết từ thống kê trình bày ở chương trước (xem 3.1.3.2). Để đánh giá hiệu quả thực thi của các phương pháp nói trên, chúng ta dựa trên hai tiêu chuẩn là độ phủ (recall) và độ chính xác (precision) :

- Ý nghĩa của độ phủ (recall) là tỉ lệ giữa số lần xuất hiện của từ đa nghĩa (trong ngôn ngữ nguồn) được liên kết với từ khác rỗng (trong ngôn ngữ đích) với số lần xuất hiện tổng cộng của từ đa nghĩa.

- Ý nghĩa của độ chính xác (precision) là tỉ lệ giữa số lần xuất hiện của từ đa nghĩa (trong ngôn ngữ nguồn) được liên kết đúng với số lần xuất hiện tổng cộng của từ đa nghĩa.

Từ 60,000 cặp câu trong ngữ liệu song ngữ Anh – Việt, tất cả những câu có chứa những danh từ đa nghĩa *bank, arm, interest, plant* được tập hợp lại để chuẩn bị cho việc liên kết từ (số liệu chi tiết được trình bày trong bảng 3).

	Số cặp câu		**Số cặp câu**
arm	123 / 60,000	***plant***	188 / 60,000
bank	275 / 60,000	***interest***	170 / 60,000

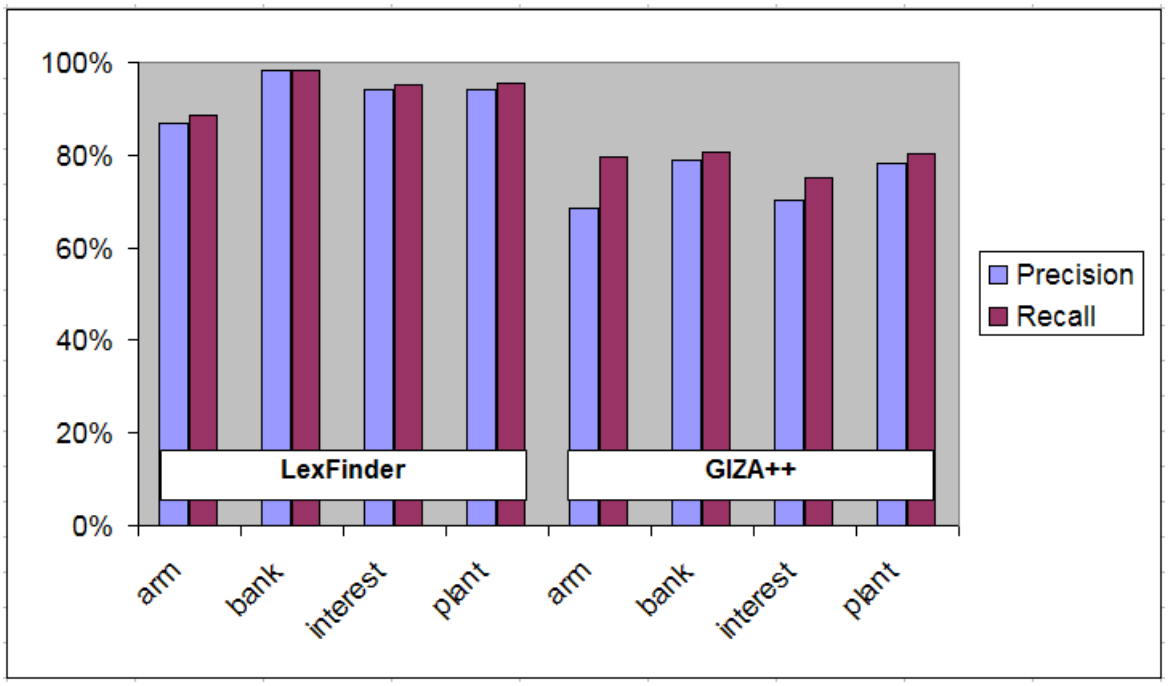
Bảng 3 – Thống kê ngữ liệu song ngữ.

Kết quả chạy thử nghiệm giải thuật đề xuất trong luận văn này (*LexFinder*) và giải thuật ứng dụng liên kết từ thống kê (xem 3.1.3.2) :

	LexFinder		***GIZA++***	
	Recall	**Precision**	**Recall**	**Precision**
Arm	0.8861	0.8699	0.7956	0.6861
Bank	0.9818	0.9818	0.8071	0.7892
Interest	0.9529	0.9411	0.7522	0.7041
Plant	0.9534	0.9418	0.8043	0.7826

Bảng 4 – Kết quả thử nghiệm của *LexFinder* và *GIZA++*.

Kết quả thực thi của hai phương pháp nói trên được kiểm định bằng tay do nhãn ngữ nghĩa của chúng không giống nhau (LexFinder dùng hệ thống nhãn ngữ nghĩa LLOCE, GIZA++ sử dụng trực tiếp các mẫu dịch từ trên ngôn ngữ đích làm nhãn ngữ nghĩa).



Hình 2 – Biểu đồ so sánh kết quả thực thi giữa *LexFinder* và *GIZA++*.

4.2 HUẤN LUYỆN MÔ HÌNH HỌC PHÂN LOẠI

Từ ngữ liệu có gán nhãn ngữ nghĩa (LLOCE) thu được từ phân đoạn xây dựng ngữ liệu nói trên, chúng ta huấn luyện mô hình phân loại ngữ nghĩa Bayes như đã trình bày ở trên (xem 3.2.2) và kiểm định chất lượng ban đầu qua bộ test độc lập được rút ra từ nhiều nguồn ngữ liệu khác nhau (Senseval-3, TWA Sense Tagged Data, Brown Corpus) :

	Số lượng		**Số lượng**
arm	500	***Plant***	500
bank	500	***interest***	500

Bảng 5 – Số liệu về ngữ liệu kiểm định.

Bộ ngữ liệu kiểm định trên được gán lại nhãn ngữ nghĩa (bằng tay) dựa theo hệ thống nhãn ngữ nghĩa LLOCE và chỉ gồm những mẫu ngữ cảnh có nhãn ngữ nghĩa được quan sát thấy trong ngữ liệu huấn luyện xây dựng ở bước trước.

	Ngữ nghĩa quan sát được khi xử lý văn bản song ngữ
arm	B41 – the arm, H230 – weapons and armanents
bank	L99 – banks, J104 – financial institutions
plant	A30 – living things, I108 – factories, milks
interest	J112 – interest on money, F228 – excitement

Bảng 6 – Ngữ nghĩa quan sát được của các từ đa nghĩa khi xử lý văn bản song ngữ.

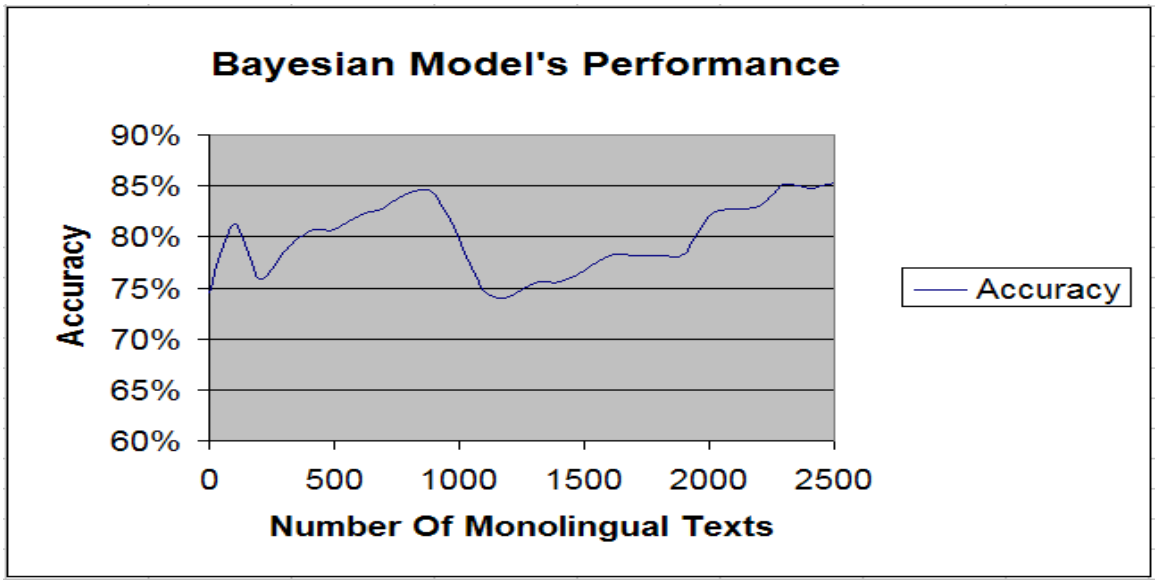
Sau khi huấn luyện mô hình phân loại Bayes và chạy thử trên bộ ngữ liệu kiểm định nói trên, ta thu được kết quả khởi đầu sau đây :

	Độ chính xác		**Độ chính xác**
Arm	0.7447	***plant***	0.7300
Bank	0.8120	***interest***	0.7000

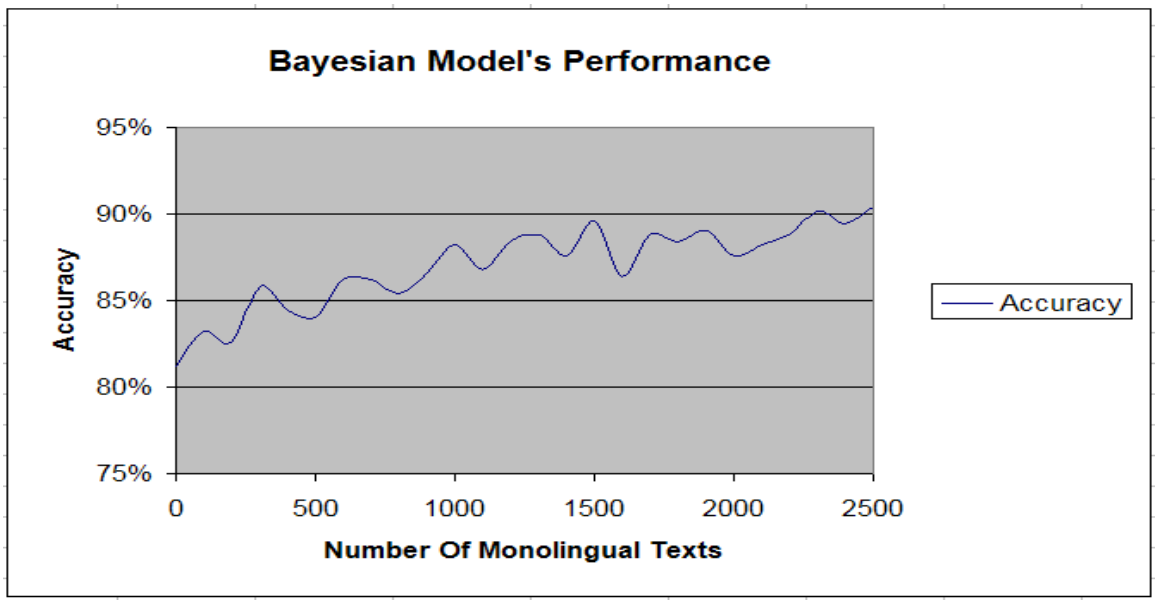
Bảng 7 – Độ chính xác của mô hình phân loại ngữ nghĩa

4.3 KHAI THÁC NGỮ LIỆU ĐƠN NGỮ

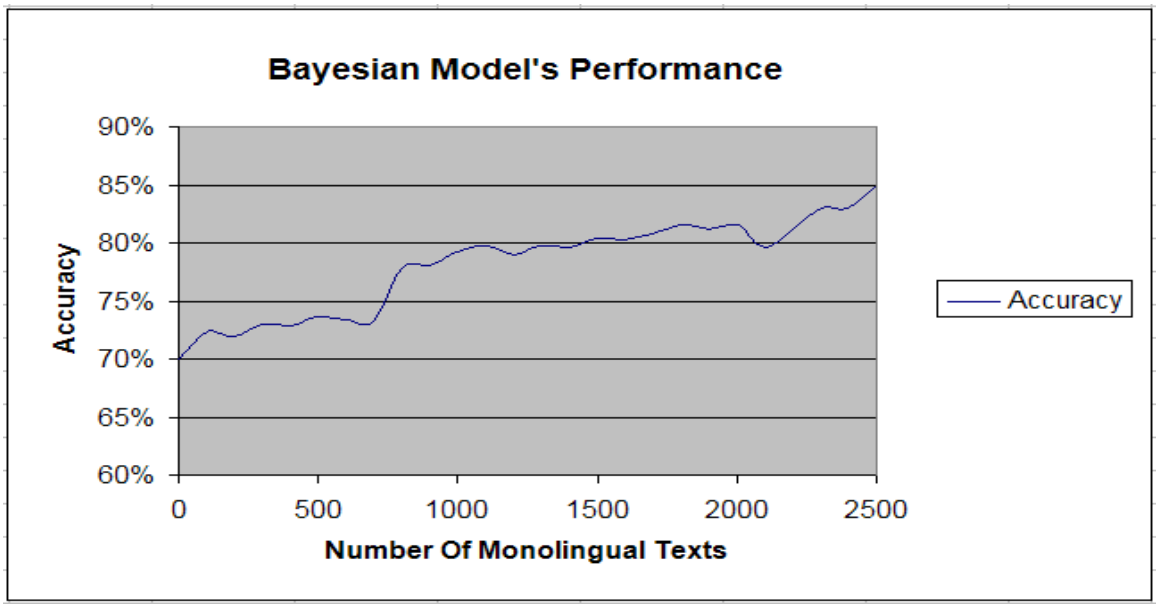
Để tăng cường độ chính xác của mô hình khử nhập nhằng ngữ nghĩa đã được huấn luyện ở các bước trên, ta tiến hành khai thác ngữ liệu đơn ngữ như mô tả ở trên (xem 3.2.3). Ngữ liệu đơn ngữ được thu thập từ ngữ liệu Wall Street Journal (WSJ), Susanne Corpus, Dow Jones Articles Archive bao gồm xấp xỉ 2,500 văn bản đơn ngữ, độc lập.



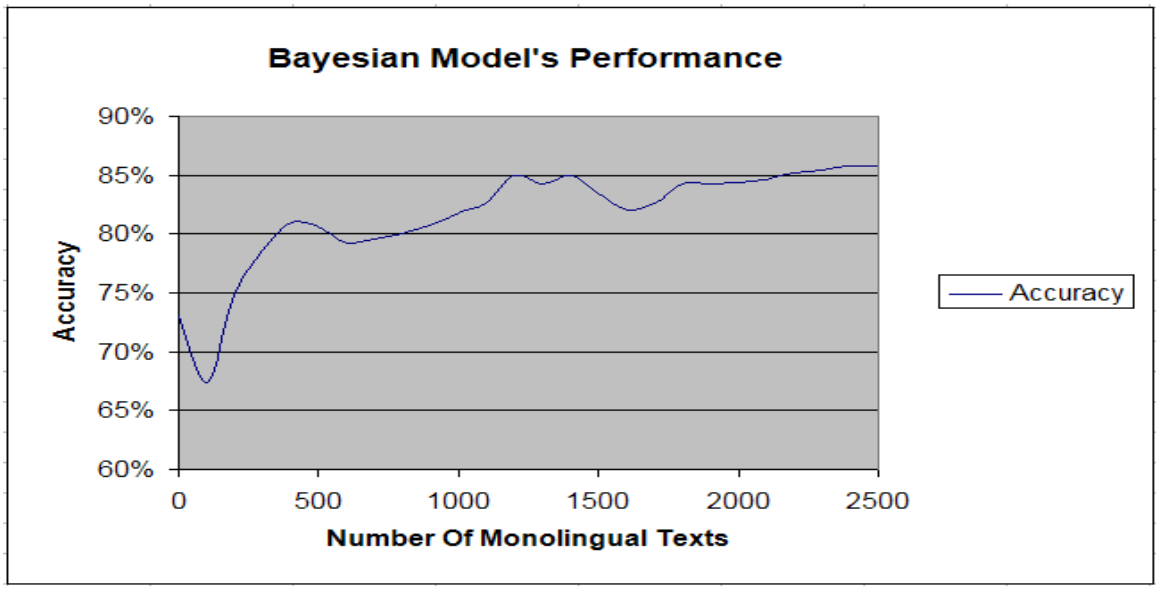
Hình 3 – Độ chính xác của mô hình khử nhập nhằng ngữ nghĩa cho từ *arm* trong quá trình khai thác ngữ liệu đơn ngữ.



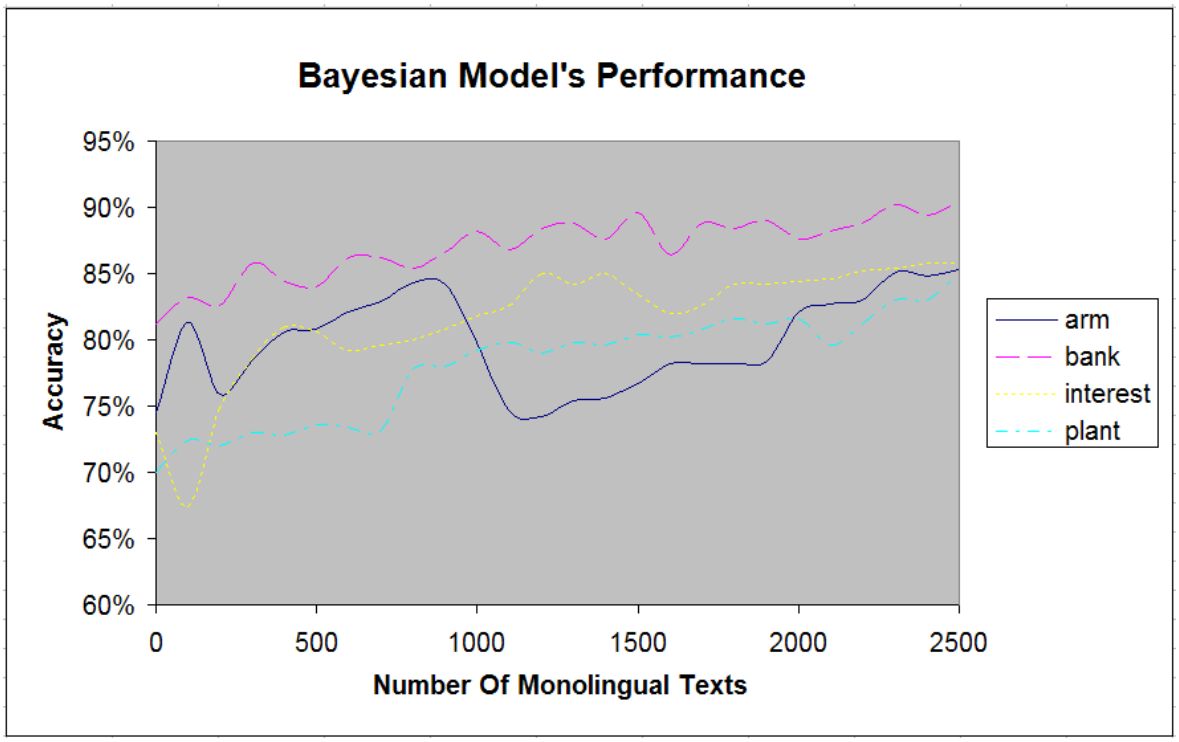
Hình 4 – Độ chính xác của mô hình khử nhập nhằng ngữ nghĩa cho từ *bank* trong quá trình khai thác ngữ liệu đơn ngữ.



Hình 5 - Độ chính xác của mô hình khử nhập nhằng ngữ nghĩa cho từ *plant* trong quá trình khai thác ngữ liệu đơn ngữ.



Hình 6 - Độ chính xác của mô hình khử nhập nhằng ngữ nghĩa cho từ *interest* trong quá trình khai thác ngữ liệu đơn ngữ.

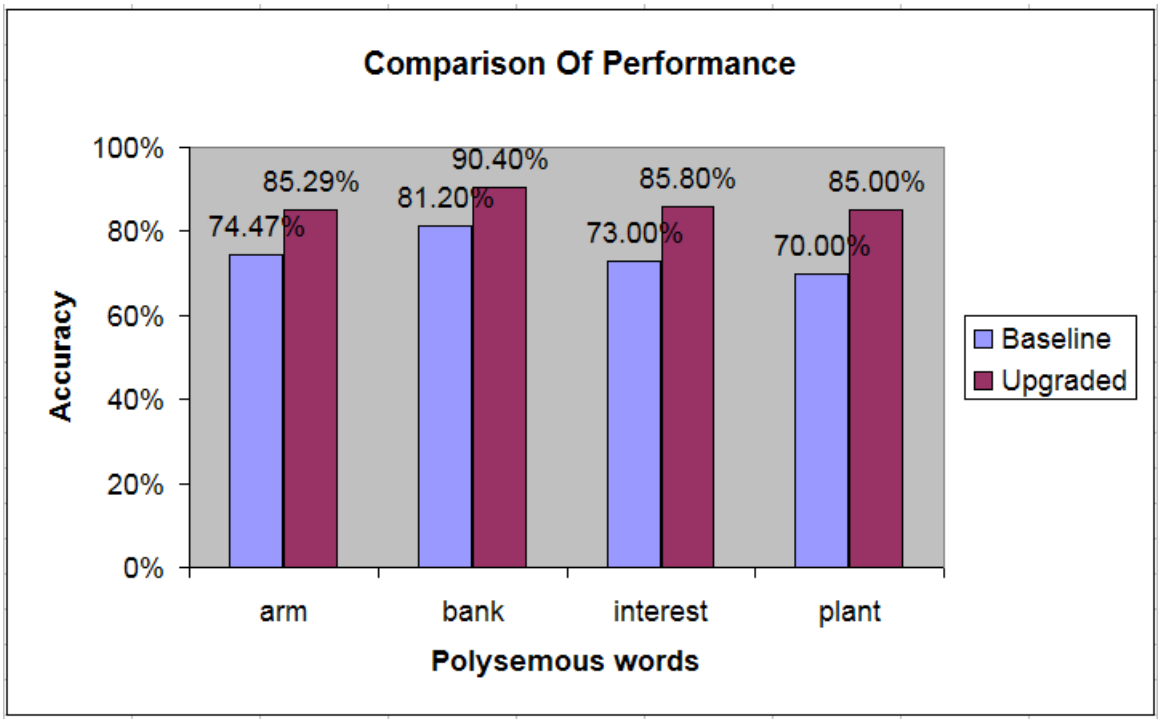


Hình 7 – Tổng hợp độ chính xác của mô hình khử nhập nhằng ngữ nghĩa cho các từ trong quá trình khai thác ngữ liệu đơn ngữ.

Sau khi kết thúc quá trình khai thác trên tập các văn bản đơn ngữ, ta kiểm định lại chất lượng của mô hình phân loại ngữ nghĩa trên tập ngữ liệu kiểm định nói trên và thu được kết quả sau :

	Độ chính xác		**Độ chính xác**
Arm	0.8529	***plant***	0.8500
Bank	0.9040	***interest***	0.8580

Bảng 8 – Độ chính xác của mô hình phân loại ngữ nghĩa sau khi được gia cố trên ngữ liệu đơn ngữ.



Hình 8 – Biểu đồ so sánh chất lượng của mô hình phân loại ngữ nghĩa trước và sau khi gia cố trên ngữ liệu đơn ngữ.

Kích thước của tập ngữ liệu huấn luyện trước và sau khi khai thác trên ngữ liệu đơn ngữ :

	Kích thước trước khi khai thác trên văn bản đơn ngữ	**Kích thước sau khi khai thác trên tập văn bản đơn ngữ**
arm	123	415
bank	275	809
plant	188	435
interest	170	456

Bảng 9 – Kích thước ngữ liệu trước và sau khi khai thác trên tập hợp các văn bản đơn ngữ.

CHƯƠNG 5 – KẾT LUẬN VÀ HƯỚNG MỞ RỘNG

5.1 KẾT LUẬN

Luận văn này trình bày một cách tiếp cận mới cho bài toán xử lý nhập nhằng ngữ nghĩa. Trong đó, từ điển đồng nghĩa Anh – Việt và ngữ liệu song ngữ được sử dụng để xây dựng ngữ liệu có gán nhãn ngữ nghĩa dùng để huấn luyện mô hình phân loại ngữ nghĩa (Bayes Classification). Ngoài ra, ngữ liệu đơn ngữ cũng được khai thác trong cách tiếp cận này nhằm nâng cao chất lượng khử nhập nhằng ngữ nghĩa.

Qua các kết quả thử nghiệm ở trên, ta nhận thấy với cách làm kết hợp các nguồn tài nguyên khác nhau gồm ngữ liệu song ngữ, ngữ liệu đơn ngữ, từ điển đồng nghĩa tiếng Việt cho hiệu quả xử lý nhập nhằng ngữ nghĩa khá tốt. Mặc dù chỉ mới được thử nghiệm trên một lớp nhỏ các từ đồng nghĩa (do còn hạn chế về mặt ngữ liệu) nhưng kết quả đạt được vẫn cho thấy đây là một hướng đi khả thi, có thể ứng dụng và triển khai được trong các nghiên cứu có liên quan đến vấn đề khử nhập nhằng ngữ nghĩa.

Việc kết hợp từ điển đồng nghĩa tiếng Việt với tập hợp các văn bản song ngữ để xây dựng ngữ liệu có gán nhãn ngữ nghĩa đã giải quyết được vấn đề phân mảnh dữ liệu thường gặp đối với các hướng tiếp cận trước đây, dùng luôn các mẫu dịch trong ngôn ngữ đích để làm nhãn ngữ nghĩa. Ví dụ, khi gán nhãn ngữ nghĩa cho một từ đa nghĩa, chẳng hạn như *bank* vốn có hai cách diễn dịch trên ngôn ngữ đích cho cùng một sắc thái ý nghĩa : *ngân hàng, nhà băng*. Với các cách tiếp cận không xét đến cơ sở đồng nghĩa trong tiếng Việt, *ngân hàng* và *nhà băng* sẽ được xem như hai nhãn ngữ nghĩa khác nhau và như thế, ngữ liệu huấn luyện thu được sẽ bị phân mảnh và sẽ ảnh hưởng đến hiệu quả thực thi.

Cuối cùng, việc khai thác ngữ liệu đơn ngữ để cải thiện chất lượng của mô hình phân loại ngữ nghĩa cũng cho kết quả thực nghiệm rất khả quan. Lợi điểm của việc mở rộng tập mẫu học trên ngữ liệu đơn ngữ là nhằm khắc phục những nhược điểm của ngữ liệu song ngữ như độ phủ thấp (do bị giới hạn trong một lĩnh vực nhất định), dữ liệu thưa, … Hơn thế nữa, nguồn tài nguyên văn bản song ngữ Anh – Việt nói riêng hiện nay cũng chưa thật sự phong phú nên việc bổ sung thêm các mẫu học bằng cách khai thác trên ngữ liệu đơn ngữ là cần thiết để xây dựng một mô hình khử nhập nhằng ngữ nghĩa với chất lượng tốt.

5.2 HƯỚNG MỞ RỘNG

Nhìn chung, bên cạnh những ưu điểm phân tích ở trên, hướng tiếp cận trong luận văn này còn một số điểm hạn chế cần phải khắc phục như sau :

- Việc khai thác mẫu học trên ngữ liệu đơn ngữ còn chưa hoàn toàn tự động (việc tập hợp các văn bản đơn ngữ vẫn do con người phụ trách). Để khắc phục nhược điểm này, chúng ta có thể kết hợp mô hình nói trên với một động cơ tìm kiếm trên các website có chứa các văn bản tiếng Việt. Theo đó, quá trình tìm kiếm sẽ được định hướng bởi tập các từ xuất hiện với xác suất cao trong các ngữ cảnh của từ đa nghĩa tương ứng.

- Ngữ liệu song ngữ dùng làm nền tảng xây dựng bộ ngữ liệu có gán nhãn ngữ nghĩa có độ phủ tương đối thấp dẫn đến tình trạng mô hình học không quan sát được hết tất cả các ngữ nghĩa có thể có của từ đa nghĩa. Để khắc phục nhược điểm này chúng ta có thể kết hợp ngữ liệu song ngữ với những kho ngữ liệu hạt giống (là những kho ngữ liệu kích thước nhỏ nhưng có độ phủ cao) để giúp cho mô hình phân loại có thể quan sát đầy đủ tất cả các ngữ nghĩa có thể có trước khi tiến hành khai thác trên ngữ liệu đơn ngữ.

- Về phương diện kỹ thuật xử lý, bộ tách từ và gán nhãn từ loại đưa vào sử dụng có tốc độ xử lý còn chậm và độ chính xác thì cũng chưa thật sự hoàn chỉnh dẫn gây ra nhiễu và làm giảm độ chính xác trong quá trình liên kết từ cho ngữ liệu song ngữ. Do đó, để có thể xây dựng một mô hình xử lý nhập nhằng ngữ nghĩa hoàn thiện hơn, chúng ta cũng cần chú trọng đến việc cải tiến những công cụ xử lý ngôn ngữ ở mức thấp như tách từ, gán nhãn từ loại, …

Một số hướng nghiên cứu trong tương lai nhằm hoàn thiện và phát triển hướng tiếp cận đề xuất trong luận văn này :

- Triển khai mô hình trên tập tất cả những từ đa nghĩa thường gặp khi dịch Anh – Việt.

- Kết hợp các nguồn tri thức ngôn ngữ khác trong xử lý nhập nhằng ngữ nghĩa để nâng cao hiệu quả thực thi.

- Tích hợp mô hình phân loại ngữ nghĩa nói trên vào hệ dịch máy Anh – Việt hiện tại (được xây dựng và phát triển bởi nhóm VCL) để nâng cao chất lượng dịch.

TÀI LIỆU THAM KHẢO

[1] Michael Lesk (1986), "*Automatic sense disambiguation using machine readable dictionaries : How to tell a pine cone from an ice cream cone*". In Proceedings of 1986 SIGDOC Conference, pages 24 – 6, Ontario, Canada.

[2] Brown, P., S.A. Pietra, V.J.D. Pietra, and R. Mercer (1991), "*Word Sense Disambiguation Using Statistical Methods*". In Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics, pages 264 - 270.

[3] Dagan, Ido, Alon Itai, Ulrike Schwall (1991), "*Two Language Are More Informative than One*". In Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics, pages 130 – 137.

[4] David Yarowsky (1992), "*Word Sense Disambiguation Using Statistical Models of Roget's Categories Trained On Large Corpora*". In Proceedings of the International Conference on Computational Linguistics, pages 454 – 460.

[5] Gale, W., K. Church, and D. Yarowsky (1992), "*Using Bilingual Materials to Develop Word Sense Disambiguation Methods*". In Proceedings of the 4th International Conference on Theoretical and Methodological Issues in Machine Translation, pages 101 - 112.

[6] Yarowsky D. (1993), "*One sense per collocation*". In Proceedings of the ARPA Human Language Technology Workshop, Princeton, NJ, pages 266 – 271.

[7] Brown et al. (1993), "*The mathematics of Machine Translation : Parameters Estimation"*. In Computational Linguistic, 19(2).

[8] Dagan I., Itai A. (1994), "*WSD using a second language monolingual corpus*". Computational Linguistic, 20(4), pages 563 – 596.

[9] Yarowsky D. (1995), "*Unsupervised word sense disambiguation rivaling supervised methods*". In Proceedings of the 33rd ACL, Cambridge, MA, USA, pages 189 – 196.

[10] Hwee N., Lee H. (1996), "*Integrating multiple knowledge sources to disambiguate word senses: an exemplar-based approach*". In Proceedings of the 34th ACL, Santa Cruz, CA, USA, pages 40 – 47.

[11] Mooney R. J. (1996), "*Comparative Experiments on Disambiguating Word Senses : An Illustration of the Role of Bias in Machine Learning*". In Proceedings of EMNLP096.

[12] Phillip Resnik, David Yarowsky (1997), "*A perspective on word sense disambiguation methods and their evaluation*". In Proceedings of the ACL SIGLEX Workshop on Tagging Text with Lexical Semantics : Why, What, and How, pages 79 – 86.

[13] Ker S. J, Chang J. S. (1997), "*A Class-based Approach to Word Alignment*". Computational Linguistic, 23(2), pages 313 – 343.

[14] Pedersen, T. (2000), "*A Simple Approach to Building Ensembles of Naive Bayesian Classifiers for Word Sense Disambiguation*". In Proceedings of the North American Chapter of the Association for Computational Linguistics (NAACL), pages 63 - 69.

[15] Christopher D. Manning, Hinrich Schutze (2000), "*Foundations of Statistical Natural Language Processing*". 3rd Edition, The MIT Press, Cambridge, Massachusetts, London, England.

[16] Rada F. Mihalcea, Dan I. Moldovan (2001), "*Pattern learning and active feature selection for word sense disambiguation*". In Proceedings of the 2nd International Workshop on Evaluating Word Sense Disambiguation Systems (SENSEVAL - 2), pages 127 – 130.

[17] Diab M., Resnik P. (2002), "*An Unsupervised Method for Word Sense Tagging using Parallel Corpora*". In Proceedings of the 40th ACL, Philadelphia, USA, pages 255 – 262.

[18] Dien D. et al. (2002), "*Word alignment in English – Vietnamese bilingual corpus*". In Proceedings of EALPIIT'02, Hanoi, Vietnam (1/2002), pages 3 – 11.

[19] D.Dien, H.Kiem (2002), "*Building a training corpus for word sense disambiguation in English – Vietnamese Machine Translation*". In Proceedings of Workshop on Machine Translation in Asia, COLING – 02, pages 26 – 32.

[20] Dien D. Kiem H. (2002), "*Bilingual corpus and word sense disambiguation in the English – to – Vietnamese Machine Translation*". In Proceedings of the 1st APIS, Bangkok, Thailand, pages 8 – 15.

[21] Li C., Li H. (2002), "*Word Translation Disambiguation Using Billingual Bootstrapping*". In Proceedings of the 40th ACL, Philadelphia, USA, pages 264 – 270.

[22] Hwee Tou Ng., Bin Wang, Yee Seng Chan (2002), "*Exploiting parallel texts for Word Sense Disambiguation*". In Proceedings of the 41st Annual Meeting of the Association for Computational Linguistic (ACL – 03), pages 455 – 462, Sapporo, Japan.

[23] Radu Florian, David Yarowsky (2002), "*Modelling consensus: Classifier combination for Word Sense Disambiguation*". In Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing, pages 25 – 32.

[24] Cuong Anh Le, Akira Shimazu (2004), "*High WSD Accuracy using Naïve Bayesian classifier with rich features*". In Proceedings of PACLIC 18, Waseda University, Tokyo.

[25] David Vickrey (2005), "*Word Sense Disambiguation for Machine Translation*". In Proceedings of HLT/EMNLP 2005.

[26] Dinh Dien (2006), "Natural Language Processing". The VNU Press, Ho Chi Minh, Vietnam.

PHỤ LỤC

A. DANH SÁCH NHÃN NGỮ PHÁP TIẾNG ANH

STT	Nhãn	Mô tả	Ví dụ
1.	CC	Coordinating conjunction (liên từ)	and, or, but, …
2.	CD	Cardinal number (số từ)	1, 2, one, two, …
3.	DT	Determiner (định từ)	the, a, an, …
4.	EX	Existential “there” (“có”)	There
5.	FW	Foreign word (từ nước ngoài)	
6.	IN	Preposition or subordinating conjunction (giới từ)	in, on, at, …
7.	JJ	Adjective (tính từ)	big, good, hard, …
8.	JJR	Adjective, comparative (tính từ so sánh hơn)	bigger, better, …
9.	JJS	Adjective, superlative (tính từ so sánh cực cấp)	biggest, best, …
10.	LS	List item marker (dấu liệt kê)	:
11.	MD	Modal (từ tình thái)	can, may, might
12.	NN	Noun, singular / mass (danh từ số ít, không đếm được)	book, sugar, action
13.	NNS	Noun, plural (danh từ số nhiều)	books, children
14.	NP	Proper noun, singular (danh từ riêng số ít)	John, Hanoi
15.	NPS	Proper noun, plural (danh từ riêng số nhiều)	IBMs, Fords, ...
16.	PDT	Pre-determiner (tiền chỉ định từ)	this, each, some ...
17.	POS	Possesive ending (dấu cuối của sở hữu cách)	‘s
18.	PP	Personal pronoun (đại từ nhân xưng)	I, you, he

19.	PP$	Possesive pronoun (đại từ sở hữu)	mine, yours, his
20.	RB	Adverb (trạng từ)	slow, hardly
21.	RBR	Adverb, comparative (trạng từ so sánh hơn)	slower, faster
22.	RBS	Adverb, superlative (trạng từ so sánh cực cấp)	slowest, fastest
23.	RP	Particle (tiểu từ)	on, off
24.	SYM	Symbol (ký hiệu)	
25.	TO	"to" (từ "to")	
26.	UH	Interjection (thán từ)	oh !
27.	VB	Verb, base form (động từ nguyên thể)	work, write
28.	VBD	Verb, past tense (động từ quá khứ)	worked, wrote
29.	VBG	Verb, gerund or present participle (danh động từ / hiện phân từ)	working, writing
30.	VBN	Verb, past participle (động từ quá khứ, phân từ)	worked, written
31.	VBP	Verb, non 3rd person singular present (động từ không phải ngôi thứ 3 số ít hiện tại)	work, write
32.	VBZ	Verb, 3rd person singular present (động từ ngôi thứ 3, số ít hiện tại)	works, writes
33.	WDT	Wh-determiner (định từ bắt đầu bằng Wh)	which, what
34.	WP	Wh-pronoun (đại từ bắt đầu bằng Wh)	who, where
35.	WP$	Possessive Wh-pronoun (đại từ sở hữu bắt đầu bằng Wh)	whose
36.	WRB	Wh-adverb (trạng từ bắt đầu bằng Wh)	when, where

B. HỆ THỐNG NHÃN NGỮ NGHĨA LLOCE

LLOCE được phân thành 3 cấp : cấp 1 gồm 14 chủ đề, cấp 2 gồm 129 nhóm, cấp 3 gồm 2449 lớp ngữ nghĩa với tổng số 16,000 mục từ với khoảng 25,000 ngữ nghĩa. Mỗi lớp ngữ nghĩa gồm các từ đồng nghĩa hoặc có quan hệ ngữ nghĩa với nhau và mang nhãn ngữ nghĩa là tên lớp đó.

Chủ đề	**Mô tả**
A	Life and living things – Sự sống và các sinh vật
B	The Body : its Functions and Welfare – Cơ thể, chức năng và việc chăm sóc
C	People and the Family – Con người và gia đình
D	Buildings, Houses, Clothes, Belongings, Personal Care – Công trình xây dựng, nhà cửa quần áo, đồ đạc và tiện nghi cá nhân
E	Food, Drink and Farming – Thực phẩm, đồ uống và nghề nông
F	Feelings, Emotions, Attributes and Sensations – Cảm xúc, xúc cảm, thái độ và cảm giác
G	Thought and Communication, Language and Grammar – Tư duy và thông tin, ngôn ngữ và văn phạm
H	Substances, Material, Objects and Equipment – Chất liệu, vật liệu, đồ vật và trang thiết bị
I	Arts and Craft, Science and Technology, Industry and Education – Nghệ thuật và nghề thủ công, khoa học và công nghệ, công nghiệp, giáo dục
J	Numbers, Measurement, Money and Commerce – Số, đo lường, tiền tệ và thương mại
K	Entertainment, Sports and Games – Giải trí, Thể thao và các môn thi đấu
L	Space and Time – Không gian và Thời gian
M	Movement, Location, Travel and Transport – Dịch chuyển, vị trí, du hành và vận tải
N	General and Abstract Terms – Các thuật ngữ khái quát và trừu tượng

C. MỘT SỐ CÔNG CỤ XỬ LÝ NGÔN NGỮ TỰ NHIÊN

C.1 BRILL TAGGER

Đây là công cụ gán nhãn ngữ pháp (Part Of Speech) cho tất cả các từ trong một câu hay trong một văn bản tiếng Anh. Trong đó, hệ thống nhãn ngữ pháp được sử dụng là Penn Tree Bank. Công cụ này được xây dựng và phát triển bởi Eric Brill (Eric Brill, PhD Thesis, 1993). Nền tảng của phương pháp này là dựa trên tập luật ngữ cảnh được rút ra trong quá trình huấn luyện trên ngữ liệu có gán nhãn ngữ pháp.

Công cụ này hiện cho phép người sử dụng download miễn phí tại địa chỉ http://research.microsoft.com/en-us/um/people/brill/

C.2 VCL WORD SEGMENTATION

Đây là công cụ tách từ tiếng Việt (do trong tiếng Việt, ranh giới giữa các từ trong cùng một câu vốn không rõ ràng) được xây dựng và phát triển bởi các tác giả Hoàng Công Duy Vũ, Nguyễn Lê Nguyên (thuộc nhóm VCL). Công cụ được release dưới dạng file .dll (libVCL_VieWoSeg.dll) bao gồm các chức năng chính sau :

- MMSVM_Segment(CString sInput,int nType = 0) : thực hiện tách từ cho 1 câu đầu vào, kết quả sẽ là 1 câu đã tách từ.

- MMSVM_Segment_File(CString fileIn,CString fileOut,int nType = 0) : thực hiện tách từ cho 1 file text đầu vào, kết quả sẽ là 1 file text đã tách từ.

- MMSVM_Segment2List(CString fileIn,CString fileOut,int nType = 0) : thực hiện tách từ cho 1 thư mục chứa các file text đầu vào, kết quả sẽ là 1 thư mục chứa các file text đã tách từ.

C.3 GIZA++

Đây là công cụ thực hiện mô hình dịch thống kê (SMT). Công cụ này cho phép chúng ta liên kết từ trong ngữ liệu song ngữ (xác suất liên kết), xác suất chuyển vị trí (trật tự từ) và xác suất dịch từ tương ứng. Công cụ này nằm trong bộ công cụ EGYPT (http://www.clsp.jhu.edu/ws99/projects/mt) của trường Đại học John Hopkins, Mỹ. Có thể sử dụng bộ công cụ này để dịch thống kê giữa hai ngôn ngữ bất kỳ.

D. CÁC KHO NGỮ LIỆU ĐƠN NGỮ TIẾNG ANH

D.1 NGỮ LIỆU TIẾNG ANH PENN TREE BANK (PTB)

Đây là kho ngữ liệu tiếng Anh thông dụng nhất gồm 4.5 triệu từ, được rút từ các kho ngữ liệu là các báo WSJ (Wall Street Journal) và được gán nhãn từ loại cùng với phân tích cây cú pháp. Kho PTB có hai dạng : dạng do máy thực hiện tự động không chính xác hoàn toàn (có thể download miễn phí) và dạng có người hiệu chỉnh được bán với giá cao (liên hệ ACL/DCI). Đây là kho ngữ liệu mà hầu hết các chương trình gán nhãn từ loại hay cú pháp sử dụng để huấn luyện hay đánh giá.

Địa chỉ liên hệ : http://www.ldc.upenn.edu/Catalog/LDC2000T43.html

D.2 NGỮ LIỆU TIẾNG ANH SUSANNE

Đây là ngữ liệu tiếng Anh được xây dựng bởi một nhóm các nhà ngôn ngữ học máy tính (đứng đầu là Geoffrey Sampson) thuộc đại học Sussex, England. SUSANNE (Surface and Underlying Structural Analyses of Naturalistic English) đánh dấu tiếng Anh về hình thái, từ loại, cú pháp, … Nó gồm 128,000 từ chia làm bốn lĩnh vực : A – Báo chí, G – Thư từ, J – Khoa học, N – Truyện được rút từ ngữ liệu Brown (khoảng 1 triệu từ). Để mua kho ngữ liệu này, chúng ta có thể liên hệ đại học Oxford, England.

Địa chỉ liên quan :
http://www.essex.ac.uk/linguistics/clmt/w3c/corpus_ling/content/corpora/list/public/susanne.html

D.3 NGỮ LIỆU TIẾNG ANH SEMCOR

Đây là ngữ liệu gồm 250,000 mục từ được rút từ kho ngữ liệu Brown (gồm 1,000,000 từ). Danh từ trong ngữ liệu SEMCOR được gán nhãn ngữ nghĩa theo nhãn là số thứ tự của synset tương ứng trong hệ thống nhãn ngữ nghĩa WordNet. Ví dụ : từ "*plant*" mang nhiều nghĩa, mỗi nghĩa sẽ nằm trong một synset, chẳng hạn : synset_1 là "*nhà máy*", synset_2 là "*thực vật*". Khi đó, nhãn synset của từ "*plant*" trong ngữ liệu SEMCOR sẽ là 1 hoặc 2, tùy theo nghĩa thực tế của ngữ cảnh.

SEMCOR có thể download miễn phí với mục đích nghiên cứu tại địa chỉ sau http://www.cse.unt.edu/~rada/downloads.html

D.4 NGỮ LIỆU TIẾNG ANH SEMCOR

Đây là kho ngữ liệu gồm 1,000,000 từ được xây dựng từ tập hợp các văn bản, sách, tài liệu xuất bản ở Mỹ (1961). Nó bao gồm 500 mẫu tài liệu (mỗi mẫu khoảng 2000 từ) trích từ nhiều nguồn, nhiều lĩnh vực (khoa học, tin tức, báo chí, …). Ngữ liệu

Brown hiện được cung cấp miễn phí cho mục đích nghiên cứu tại ICAME (International Computer Archive Of Modern and Medieval English).

Địa chỉ liên quan : http://nora.hd.uib.no/whatis.html