Contents

[CHƯƠNG 1 : MỞ ĐẦU 3](#_Toc300386586)

[CHƯƠNG 2 : TỔNG QUAN 4](#_Toc300386587)

[2.1 GIỚI THIỆU BÀI TOÁN GÁN NHÃN NGỮ NGHĨA 4](#_Toc300386588)

[2.1.1 PHÁT BIỂU BÀI TOÁN 4](#_Toc300386589)

[2.1.2 Tầm quan trọng 4](#_Toc300386590)

[2.2 LỊCH SỬ QUÁ TRÌNH NGHIÊN CỨU 5](#_Toc300386591)

[2.2.1 TRÊN THẾ GIỚI 5](#_Toc300386592)

[2.2.2 VIỆT NAM 6](#_Toc300386593)

[2.3 Những khó khăn, thử thách 6](#_Toc300386594)

[2.4 Các hướng tiếp cận 6](#_Toc300386595)

[2.4.1 Cách tiếp cận theo luật 6](#_Toc300386596)

[2.4.2 Cách tiếp cận theo hướng ngữ liệu 7](#_Toc300386597)

[CHƯƠNG 3 : CƠ SỞ LÝ THUYẾT 9](#_Toc300386598)

[3.1 Cơ sở lý thuyết ngôn ngữ học 9](#_Toc300386599)

[3.1.1 Khái niệm về nhãn ngữ nghĩa của từ 9](#_Toc300386600)

[3.1.2 Một số hệ thống nhãn ngữ nghĩa 9](#_Toc300386601)

[3.1.2.1 Hệ thống nhãn ngữ nghĩa LLOCE 9](#_Toc300386602)

[3.1.2.2 Hệ thống nhãn ngữ nghĩa LDOCE 9](#_Toc300386603)

[3.1.2.3 Hệ thống nhãn ngữ nghĩa WordNet 9](#_Toc300386604)

[3.1.2.4 … 9](#_Toc300386605)

[3.1.3 Các nguồn tri thức cho gán nhãn ngữ nghĩa 9](#_Toc300386606)

[3.1.3.1 Thông tin hình thái 9](#_Toc300386607)

[3.1.3.2 Thông tin ngữ pháp 9](#_Toc300386608)

[3.1.3.3 Thông tin ngữ nghĩa 9](#_Toc300386609)

[3.1.4 Mạng ngữ nghĩa WordNet 9](#_Toc300386610)

[3.1.4.1 Giới thiệu WordNet 9](#_Toc300386611)

[3.1.4.2 Các loại quan hệ trong WordNet 9](#_Toc300386612)

[3.1.4.2.1 Quan hệ đồng nghĩa 9](#_Toc300386613)

[3.1.4.2.2 Quan hệ trái nghĩa 9](#_Toc300386614)

[3.1.4.2.3 Quan hệ cấp bậc 9](#_Toc300386615)

[3.1.4.2.4 Quan hệ bộ phận – toàn thể 9](#_Toc300386616)

[3.1.5 Ngữ liệu ngữ nghĩa SemCor 9](#_Toc300386617)

[3.1.5.1 Giới thiệu SemCor 9](#_Toc300386618)

[3.1.5.2 Thông số thống kê SemCor 9](#_Toc300386619)

[3.2 Cơ sở tin học 9](#_Toc300386620)

[3.2.1 Xây dựng kho ngữ liệu 9](#_Toc300386621)

[3.2.1.1 Tiêu chí chọn ngữ liệu 9](#_Toc300386622)

[3.2.1.2 Các nguồn ngữ liệu thô 9](#_Toc300386623)

[3.2.2 Chuẩn hóa ngữ liệu 9](#_Toc300386624)

[3.2.2.1 Tiền xử lý 9](#_Toc300386625)

[3.2.2.2 Gán nhãn từ loại 9](#_Toc300386626)

[3.2.2.3 Gán nhãn ngữ nghĩa 10](#_Toc300386627)

[CHƯƠNG 4 : MÔ HÌNH THUẬT TOÁN 11](#_Toc300386628)

[4.1 Mô hình đề xuất cho xây dựng bộ ngữ liệu ngữ nghĩa 11](#_Toc300386629)

[4.1.1 Dịch từ ngữ liệu SemCor 11](#_Toc300386630)

[4.1.2 Sử dụng ngữ liệu song ngữu Anh – Việt 11](#_Toc300386631)

[4.2 Tìm kiếm ngữ liệu song ngữ 11](#_Toc300386632)

[4.2.1 Xây dựng dữ liệu thô 11](#_Toc300386633)

[4.2.2 Tách từ 11](#_Toc300386634)

[4.2.3 Gán nhãn từ loại (Pos Tagger) 11](#_Toc300386635)

[4.2.4 Liên kết từ 11](#_Toc300386636)

[4.2.5 Gán nhãn ngữ nghĩa cho tiếng Anh 11](#_Toc300386637)

[4.2.6 Xây dựng bộ nhãn ngữ nghĩa 11](#_Toc300386638)

[CHƯƠNG 5 : THIẾT KẾ CÀI ĐẶT 12](#_Toc300386639)

[CHƯƠNG 6 : KẾT QUẢ - ĐÁNH GIÁ 13](#_Toc300386640)

# MỞ ĐẦU

# TỔNG QUAN

## GIỚI THIỆU BÀI TOÁN GÁN NHÃN NGỮ NGHĨA

### PHÁT BIỂU BÀI TOÁN

Trong hầu hết các ngôn ngữ, có rất nhiều từ mang những sắc thái ý nghĩa khác nhau nếu đặt trong các ngữ cảnh khác nhau. Ngữ nghĩa của một từ phụ thuộc vào ngữ cảnh xuất hiện của từ đó. Từ thực tế này, nhu cầu gán nhãn ngữ nghĩa được đặt ra với mục đích tìm kiếm một cơ chế tự động, dựa trên ngữ cảnh xung quanh một từ đa nghĩa và chỉ ra được ý nghĩa phù hợp nhất của nó.

Bài toán có thể được mô tả như sau:

* Input: Một chuỗi các từ và tập nhãn từ loại (Ví dụ như chuỗi các từ “Book that flight,” và tập nhãn Penn TreeBank của tiếng Anh.).
* Output: Một nhãn ngữ nghĩa tốt nhất cho từng từ trong chuỗi đã được đưa ra (Ví dụ: đối với chuỗi từ “Book that flight.”, nhãn ngữ nghĩa thích hợp tương ứng cho từng từ sẽ là Book…….).

Từ bài toán này, ta xây dựng nên hai bài toán con như sau:

* Xây dựng bộ ngữ liệu ngữ nghĩa cho tiếng Việt, tương tự bộ ngữ liệu ngữ nghĩa tiếng Anh (SemCor)[tltk].
* Áp dụng học có giám sát trên bộ ngữ liệu vừa thu được. Từ đó đưa ra sử dụng trên một câu tiếng Việt bất kì.

### VAI TRÒ – TẦM QUAN TRỌNG

Có thể xem việc xác định đúng nghĩa cử từ là một vấn đề trung tâm của mọi hệ xử lý ngôn ngữ tự nhiên. Hiệu quả làm việc của bộ phận xác định nghĩa của từ có một ảnh hưởng rất lớn đến chất lượng thực hiện của một hệ xử lý ngôn ngữ tự nhiên.

Rõ ràng, để gán ngữ nghĩa cho một từ, ta cần phải phân tích ngữ cảnh của nó. Vấn đề đặt ra ở đây là làm sao lập trình cho máy tính hiểu được ngữ cảnh đó. Đây chính là một trong hai bài toan cơ bản, đầy thử thách của xử lý ngôn ngữ tự nhiên. Về bản chất, gán nhãn 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, đây lại là một bước cần thiết không thể bỏ qua trong các ứng dụng liên quan đến ngôn ngữ tự nhiên:

* Rút trích thông tin tự động: Với yêu cầu thu thập các văn bản liên quan đến một vấn đề cụ thể nào đó, chúng ta cần biết nghĩa của những từ khóa sử dụng để tìm kiếm. Bài toán đặt ra ở đây là cần phải có sự khử nhập nhằng, gán ngữ nghĩa cho các từ này để hạn chế rút trích ra những thông tin không liên quan đến chủ đề mà ta quan tâm. Ví dụ: khi ta sử dụng từ khóa “đường, ngọt” để tìm kiếm thì hệ thống cần phải hiểu được rằng “đường” ở đây mang sắc thái ý nghĩa “đường ăn” chứ không phải là công trình giao thông - “đường đi”.
* Xử lý văn bản tự động: Khi hệ thống tiến hành kiểm tra lỗi chính tả, ngữ pháp thì việc gán nhãn ngữ nghĩa cũng đóng vai trò quan trọng. Ví dụ: Khi hệ thống chuẩn hóa văn bản thô, “CON TRAI ÔNG CỐC LÊN DỐC BẮN CÒ” và “CÁI CỐC BỊ VỠ” sẽ được chuẩn hóa tương ứng thành “Con trai ông Cốc lên dốc bắn cò” và “Cái cốc bị vỡ”.
* Dịch máy tự động: Gán ngữ nghĩa cho từ là một vấn đề thiết yếu của những hệ thống dịch tự động do các từ đa nghĩa trong ngôn ngữ nguồn có thể được dịch sang nhiều cách khác nhau ở ngôn ngữ đích. 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.
* Tổng hợp và nhận dạng tiếng nói: Để máy có thể phát âm chính xác thì việc gán nhãn ngữ nghĩa là hết sức cần thiết. Đối với nhận dạng tiếng nói, việc gán ngữ nghĩa cũng góp phần trong việc phân đoạn từ và phân biệt từ đồng âm khác nghĩa.

## LỊCH SỬ QUÁ TRÌNH NGHIÊN CỨU

### TRÊN THẾ GIỚI

Xử lý nhập nhằng - gãn nhãn 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. Trong một công trình có từ năm 1949, Weaver thảo luận sự cần thiết phải xác định nghĩa đúng của từ và định ra những bước cơ bản trong khử nhập nhằng – gán nhãn ngữ nghĩa (Xem thêm trong [13]). Ông nói rằng: Nếu một ai đó xem xét từng từ một trong một quyển sách thì rõ ràng người đó không thể xác định được nghĩa của tất cả các từ. Ông cũng cho rằng nếu như mở rộng vùng xem xét xung quanh từ đó thì không những xác định được nghĩa của từ đó mà còn có thể xác định thêm được nghĩa của những từ xung quanh nữa. Nhưng vùng xung quanh đó có kích thước cửa sổ xem xét là bao nhiêu ?

Năm 1950, một thí nghiệm nổi tiếng do Kaplan thực hiện nhằm tìm câu trả lời cho câu hỏi nêu trên. Kaplan dùng 7 từ để xem xét, và vùng cửa sổ xung quanh xem xét được thay đổi từ một đến hai từ mỗi bên của từ cần xem xét. Kaplan quan sát rằng độ phân giải nghĩa được đưa 2 từ trên mỗi bên của từ xem xét không tốt hơn cũng như không tệ hơn khi đưa toàn bộ câu.

Tuy vậy, trong giới chuyên môn vẫn 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 sau đó đều bị bỏ dở.

Mặc dù vậy, trong khoảng thời gian đó biểu diễn tri thức lại có những bước tiến đáng kể, đáng chú ý nhất 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 gán nhãn 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, gán nhãn ngữ nghĩa hiện là bài toán trọng tâm thu hút rất nhiều các công trình nghiên cứu.

## NHỮNG KHÓ KHĂN, THỬ THÁCH

Bài toán gãn nhãn ngữ nghĩa ở Việt Nam vẫn còn đang thực hiện ở mức khiêm tốn, do những khó khăn thử thách trong việc xử lý nhập nhằng ngữ nghĩa bao gồm:

* Ngữ nghĩa của một từ vẫn chưa có một định nghĩa rõ ràng. Để định nghĩa các sắc thái ý nghĩa của một từ, ta thường tiếp cận theo hướng từ điển. Tuy vậy, 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 ranh giới rõ ràng để phân biệt, ví dụ như từ *đi* có các sắc thái ý nghĩa sau:
* Dịch chuyển bằng hai chi dưới. Ví dụ như: tôi *đi* rất nhanh nhưng vẫn không đuổi kịp anh ấy.
* Một người nào đó đã chết. Ví dụ như: Anh ấy ra *đi* mà không kịp nói lời nào.

Hai sắc thái ý nghĩa (1) và (2) thật sự khó phân biệt nếu không 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 dùng cho gán nhãn ngữ nghĩa còn quá ít. Đây là một trong những vấn đề chính còn tồn tại, khiến cho việc tiếp cận bài toán gán nhãn ngữ nghĩa còn nhiều khó khăn. Vấn đề gán nhãn 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.
* Bên cạnh đó, ngoài kiến thức tin học, công việc xử lý ngữ nghĩa cũng cần phải có những kiến thức về ngôn ngữ học. Những kiến thức này hỗ trợ cho việc tìm mối quan hệ giữa cú pháp và ngữ nghĩa, mối quan hệ giữa các nghĩa, sự phân loại…

Những khó khăn trên đã đem đến rất nhiều trở ngại cho bài toán gán nhãn ngữ nghĩa. Cho đến nay, có thể nói vẫn chưa có 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 quan trọng và phần nào giải quyết được bài toán khử nhập nhằng ngữ nghĩa (dù chưa hoàn chỉnh). Trước tình hình phát triển hiện nay của công nghệ thông tin, chúng ta hoàn toàn có quyền hi vọng là những khó khăn nói trên sẽ dần dần được giải quyết để xây dựng mô hình xử lý gán nhãn ngữ nghĩa.

## CÁC HƯỚNG TIẾP CẬN

### 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, gán nhãn 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 gán nhãn ngữ nghĩa   
được xây dựng trên luật sẽ có nguy cơ bị sụp đổ bởi chính sức nặng của chúng.

### CÁCH TIẾP CẬN THEO 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 xử lý gán nhãn 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.

Nói tóm lại, trong các hƣớng 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. 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, ngƣời 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.

# CƠ SỞ LÝ THUYẾT

## Cơ sở lý thuyết ngôn ngữ học

### Khái niệm về nhãn ngữ nghĩa của từ

### Một số hệ thống nhãn ngữ nghĩa

#### Hệ thống nhãn ngữ nghĩa LLOCE

#### Hệ thống nhãn ngữ nghĩa LDOCE

#### Hệ thống nhãn ngữ nghĩa WordNet

#### …

### Các nguồn tri thức cho gán nhãn ngữ nghĩa

#### Thông tin hình thái

#### Thông tin ngữ pháp

#### Thông tin ngữ nghĩa

### Mạng ngữ nghĩa WordNet

#### Giới thiệu WordNet

#### Các loại quan hệ trong WordNet

##### Quan hệ đồng nghĩa

##### Quan hệ trái nghĩa

##### Quan hệ cấp bậc

##### Quan hệ bộ phận – toàn thể

### Ngữ liệu ngữ nghĩa SemCor

#### Giới thiệu SemCor

#### Thông số thống kê SemCor

## Cơ sở tin học

### Xây dựng kho ngữ liệu

#### Tiêu chí chọn ngữ liệu

#### Các nguồn ngữ liệu thô

### Chuẩn hóa ngữ liệu

#### Tiền xử lý

#### Gán nhãn từ loại

#### Gán nhãn ngữ nghĩa

# MÔ HÌNH THUẬT TOÁN

## Mô hình đề xuất cho xây dựng bộ ngữ liệu ngữ nghĩa

### Dịch từ ngữ liệu SemCor

### Sử dụng ngữ liệu song ngữu Anh – Việt

## Tìm kiếm ngữ liệu song ngữ

### Xây dựng dữ liệu thô

### Tách từ

### Gán nhãn từ loại (Pos Tagger)

### Liên kết từ

### Gán nhãn ngữ nghĩa cho tiếng Anh

### Xây dựng bộ nhãn ngữ nghĩa

Nhãn ngữ nghĩa

Phương pháp xây dựng bộ nhãn ngữ nghĩa

Tổ chức bộ nhãn ngữ nghĩa theo WordNet

# THIẾT KẾ CÀI ĐẶT

# KẾT QUẢ - ĐÁNH GIÁ