Bài toán “xác định sự tương đồng ngữ nghĩa của một cặp câu văn” là một trong những bài toán cơ bản và đã được phát triển rất sớm trong thế giới công nghệ. Hiện nay, một trong những ứng dụng phổ biến nhất của bài toán này đó là trong các hệ thống tìm kiếm văn bản dựa vào từ khóa được đưa vào ví dụ như Google Search, Yahoo Search,... Bài toán nhận vào hai câu văn và cho ra kết quả phân loại xem cặp câu đó có tương đồng với nhau về mặt ngữ nghĩa hay không? Đối với bài toán này, phương pháp truyền thống (base-line) để giải quyết là dựa trên các từ giống hệt nhau của hai câu văn. Nhưng trong khóa luận, chúng tôi sẽ sử dụng các phương pháp cao cấp hơn để có thể tìm kiếm cả những cặp từ không giống nhau nhưng đồng nghĩa như “có” - “sở hữu”,... để đánh giá. Trong khóa luận này, chúng tôi sẽ giới thiệu hệ thống “xác định sự tương đồng ngữ nghĩa trong câu văn Tiếng Việt” sử dụng phương pháp dựa trên tập dữ liệu lớn (corpus-based). Nội dung của khóa luận sẽ trình bày quá trình nghiên cứu và phát triển hệ thống bao gồm các phần: Giới thiệu về bài toán, trình bày các hướng tiếp cận hiện tại, cài đặt hệ thống, thực nghiệm hệ thống trên tập dữ liệu kiểm thử tự tạo. Kết quả thực nghiệm cho thấy, hệ thống đã đạt được những thành công nhất định trong bài toán này với điểm F-score cao nhất lên tới 77% cao hơn khá nhiều so với phương pháp truyền thống với F-score chỉ đạt 70%.