氏　　名	Raul Ernesto MENENDEZ MORA
学位（専攻分野）	博士（情報学）
学位記番号	総研大甲第1515号
学位授与の日付	平成２４年３月２３日
学位授与の要件	複合科学研究科　情報学専攻 学位規則第６条第１項該当
学位論文題目	Improving Semantic Similarity Measures for Word Pair Comparison
論文審査委員	主　査　准教授　市瀬　龍太郎 教授　山田　誠二 教授　武田　英明 客員教授　相澤　彰子 教授　山口　高平　慶應義塾大学

論文内容の要旨

The semantic web provides a common framework that allows data to be shared and reused across application, enterprise, and community boundaries. The technologies of this web of data can be used in a variety of application areas; for example: data integration, knowledge representation and analysis, cataloging services, improving search algorithms and methods, social networks, etc. In order to achieve the goals of the semantic web, it has to be able to define and to describe the relations among data (i.e., resources) on the Web.

Ontology is a formal, explicit specification of a shared conceptualization. It renders shared vocabulary and taxonomy, which models a domain with the definition of entities and/or concepts, and their properties and relations. They can be used to reason about the entities within that domain. Ontologies are one of the formal representations for organizing information in the semantic web and they are also used in artificial intelligence, systems engineering, software engineering, biomedical informatics, library science, enterprise bookmarking, and information architecture as a form of knowledge representation about the world or some part of it. In the semantic web context, since many actors provide their own ontologies, ontology matching or ontology alignment has taken a critical role for helping heterogeneous resources to inter-operate.

Ontology matching tools find classes of data that are "semantically equivalent". This process determines correspondences between concepts which are called alignments. Finding those correspondences imply a similarity assessment between the involved concepts. For this reason similarity measures plays an important role in ontology matching systems.

This thesis explores an application of the semantic models to the human way of comparing words. The ability to assess similarity lies close to the core of cognition. Semantic relatedness describes the strength of the cognitive association between two concepts. For example, man and woman are very strongly related, as are monkey and banana. The concepts screwdriver and truth, however, seem to be unrelated. Other pairs of concepts often fall somewhere in between these extremes, such as book and computer or sky-rise and window. A very straightforward technique for determining the strength of relatedness between two concepts is to find the sequence of links that connects them in a semantic network. The "closer" the concepts are to one another, i.e., the shorter the path that connects them; the more strongly they are related.

Semantic similarity of words pairs is often represented by the similarity between the concepts associated with the words. Several methods have been developed to compute words similarity, most of them operating on taxonomic dictionaries like WordNet or external corpus like the Brown Corpus. However the majority of them suffer from a serious limitation. They only focus on the semantic information shared by those words, or in the semantic differences, but they have been rarely combined in a broader perspective.

The main contributions of the thesis are:

1. A novel model for semantic similarity computation. We show that a featured based

model of similarity, where semantic differences and semantic commonalities are both considered, can be applied to word pair comparison. We demonstrate the model application by obtaining 5 new semantic similarity measures.

2. Five new semantic similarity measures. After applying the Menendez-Ichise model to the traditional WordNet based semantic similarity measures we obtained five new measures. We show four of this similarity measures outperformed their classical version while the last one performed the same as its' classical version.
3. New corpus independent information content metric. We show an analysis of taxonomic properties in corpus independent metrics. The application of this analysis allowed us to obtain a new corpus independent information content metric which generated the highest value of accuracy among the corpora dependent and the corpora independent metrics we tested.

Extensive experimental results accompany this thesis. The theoretical results of the work are intended and have been tested on comparison of traditional datasets of words pairs. However, the findings are general and formal enough so that all the discussed approaches can be applied and/or generalized to the related fields.

This thesis is oriented to both researchers and practitioners in the field of the semantic similarity, as well as interested readers from neighboring fields such as ontology matching, machine learning, natural language processing, etc.

博士論文では，与えられた２つの語の類似度をどのように測ればいいのかという問題に対して取り組んでいる．新しい類似性の尺度を提案し，人間が判断する意味的な類似度との相関関係を調べることによって，提案手法の有効性を示している．

提案する枠組みでは，与えられた２つの語に対して，人間が判定するのと同様な意味的な類似度を機械が計算する．そのために，WordNet と呼ばれる語を階層的に整理したデータベースを用いて，類似度の計算を行う．その際に，語の階層の深さ，および，語同士の共通点，差異点に着目することによって，人間の意味的な類似度判定に近い類似性尺度が構築可能になるというのが，本論文の主張である．

本論文は，全５章からなる．第１章「Introduction」では，本論文の背景，目的について説明し，この論文で取り組む研究課題を，（１）意味的な共通点，差異点を考慮すると類似度の計算が改善可能か．（２）既存の類似性尺度に（１）を取り入れた場合にどのようになるか．（３）語の階層を考慮すると従来の手法を改善できるか．の，３点に整理し，この論文の成果を明らかにしている．

第２章「Related Work」では，この論文で述べられる手法の理解に必要な関連研究について述べている．まず，オントロジーマッチングを取り上げ，オントロジーの定義や関連する技術について述べている．次に，意味的な距離の基本的な定義を紹介し，研究で使用するデータベースの WordNet について述べている．そして，意味的な関連性や類似性の尺度の関連研究などについて述べ，博士論文の研究上の位置づけを明らかにしている．

第３章「A Corpus Independent Information Content Metric」では，上述の（３）の課題に取り組んでいる．最初に，WordNet を利用したノードに基づく類似性尺度の詳細について説明し，次に，従来から使われている手法の問題点を明らかにしている．そして，語の階層を考慮することで，その問題点を解決する従来手法を拡張した手法を提示し，実験的にその手法の有効性を確認している．

第４章「The Menendez-Ichise Model」では，上述の（１），（２）の課題に取り組んでいる：まず，類似性尺度をエッジに基づくものとそれ以外に分けて説明し，次に，Tversky の類似性尺度の抽象モデルに基づいて，意味の共通点と差異点を利用した類似性計算モデル，Menendez-Ichise モデルを提案している．この Menendez-Ichise モデルを WordNet に基づく類似性計算の枠組みに適用した結果，５種類の新たな類似性尺度が得られた．さらに，これらと第３章で提案した手法とを組み合わせて，実験的に，提案手法の有効性を確認している．

第５章「Conclusion and Future Work」では，博士論文の結論をまとめている．さらに，今後の展望について述べている．

上記のように，本博士論文は，語の階層の深さ，および，語同士の共通点，差異点に着目することによって，従来の２つの語の類似性尺度を改善し，人間の意味的な類似度判定により近い類似度計算が可能であることを示した点で，この研究分野の発展に貢献するものである．また，ここで取り組んだ語の類似性を計算するという研究課題は，データの意味的な統合などの具体的なタスクのみならず，人工知能，言語学などの分野において，基盤となる研究であるため，基盤技術開発という観点からも意義があると認められる．さら

に，博士論文の内容は，２本の査読付国際会議論文として発表されている他，１本の査読付ジャーナル論文で発表されており，社会からも評価されている．以上より，本論文は博士論文として，十分な水準であると審査委員全員一致で認められた．