博　士　論　文

◇知識の利用と共有

Improving Semantic Similarity Measures for Word Pair Comparison

Raul Ernesto Menendez-Mora
rmenendezm@gmail.com
総合研究大学院大学
指導教員：市瀬　龍太郎
博士（情報学），2012年3月23日　取得

キーワード：WordNet, semantic similarity measures, information content, knowledge.

概　要：The semantic web provides a common framework that allows data to be shared and reused across application, enterprise, and community boundaries. Ontologies are one of the formal representation for organizing information in the semantic web. In the semantic web context, since many actors provide their own ontologies, ontology matching 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. Finding those correspondences imply a semantic similarity assessment.

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 or external corpus. However, the majority of them 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.

In this thesis we developed and applied a model of semantic similarity computation for word pair comparison. This model considers the semantic commonalities and the semantic differences as the core of its approach. By applying the model five new WordNet-based semantic similarity measures were created. Four of this measures obtained higher values of correlation with human judgment than their original expressions, while the fifth one remained as competitive as their original version.

We also studied WordNet taxonomic properties to extend a corpus-independent information content metric. The application of this new metric in one of the previously developed node-based semantic similarity allowed us to obtain the highest value of correlation with respect to human judgment. This thesis provides a general an extensible approach of semantic similarity computation for word pair comparison.

主な公表論文：Raul Ernesto Menendez-Mora and Ryutaro Ichise: Toward simulating the human way of comparing concepts, *IEICE Trans.*, Vol. E94-D, No. 7, pp. 1419-1429 (July 2011)

現　職：Assistant Professor and Researcher at Universidad Antonio Narino, Bogota, Colombia.

論文入手先：http://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=1&cad=rja&ved=0CCgQFjAA&url=http%3A%2F%2Fwww.nii.ac.jp%2Fgraduate%2Fthesis%2Fpdf%2F201203%2Fmenendez_Dr_thesis.pdf&ei=KyZOUMadHYqi9QSSzoCwBg&usg=AFQjCNElLC-h-b3yZF3EJVXkb8wICdQEkw&sig2=tfmb9DyEhPBWuiQCYjSV0g

抱　負：My future research will concentrate on both the theoretical and practical issues of semantic similarities. However, I plan to consider them not only in the scope of the Ontology Matching but also it's application into dynamic systems for meaningful predictions in the biomedical field.

◇ Webインテリジェンス

ソーシャルタギングからのことばが指し示す実世界対象の表現獲得

馬場　雪乃
yukino_baba@mist.i.u-tokyo.ac.jp
東京大学大学院情報理工学系研究科
指導教員：本位田　真一
博士（情報理工学），2012年6月21日
取得

キーワード：webマイニング，ソーシャルタギング．

概要：本論文では，ことばが指し示す現実世界の対象物（実世界対象）を，実世界データを用いて表現するという問題に取り組んだ．ここでは，カメラやセンサなどのデバイスを通じて現実世界から直接取得できるデータのことを実世界データと呼ぶ．実世界データのうち，写真（視覚データ）と場所（地理データ）を代表例として取り上げ，「ことばが指し示す写真」，すなわちあることばの指示対象（例．「犬」）が写っている写真を獲得するという課題と，「ことばが指し示す場所」，すなわちあることばが指し示す場所を地理的な領域として獲得するという課題に取り組んだ．

獲得のためのデータ源として，Webコンテンツに対して人々がタグ（コンテンツを説明するキーワード）を付与する，ソーシャルタギングと呼ばれる仕組みに着目した．ソーシャルタギングにより生成されるデータを用いることで，訓練データ構築等の人手を掛けずにことばが指し示す実世界対象を獲得できる．

ソーシャルタギングデータには，ことばの曖昧性とノイズタグの問題がある．ことばの曖昧性については，曖昧性が解消された「ラベル」を導入しそのラベルが指し示す対象を獲得するという方法と，ことばの曖昧性を許容し指示対象を確率表現として獲得する方法をそれぞれ採用し，その実現手法を提案した．ノイズタグの問題に対しては，一つのコンテンツに与えられた複数のタグの統合，同じタグが与えられた複数のコンテンツの利用，外部データの利用といった解決方法を提案した．

主な公表論文：馬場雪乃，石川冬樹，本位田真一：Folksonomy上のタグと関連する場所の抽出，人工知能学会論文誌，Vol. 27，No. 1，pp. 1-9 (Jan. 2012)

現職：東京大学大学院情報理工学系研究科特任研究員

論文入手先：http://yukino.moo.jp

抱負：博士論文では，人間が気ままにつくったデータをうまく加工して，コンピュータを賢くするための知識に仕立て上げることに取り組みました．今後は，人間と機械をより協力させることでお互いを賢くするための仕組みや手法を研究していきたいです．