氏　　名	NGHIEM, Minh Quoc
学位(専攻分野)	博士(情報学)
学 位 記 番 号	総研大甲第 1682 号
学位授与の日付	平成26年3月20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	Semantic Enrichment of Mathematical Expressions for Mathematical Search
論 文 審 査 委 員	主　査　　准教授　宮尾 祐介 教　授　大山 敬三 教　授　高須 淳宏 准教授　市瀬 龍太郎 教　授　相澤 彰子　国立情報学研究所

(別紙様式 2)
(Separate Form 2)

論文内容の要旨
Summary of thesis contents

The issue of retrieving semantically similar mathematical expressions has received considerable attention in mathematical information access research field. Semantic search for mathematical expressions improves search accuracy by ascertaining the contextual meaning of mathematical terms as they appear in a search-able database to generate more relevant results. Semantic enrichment, a process of associating semantic tags, usually concepts, with mathematical expressions, is an important technology in a content-based search engine for mathematical expressions.

The best-known open markup format for representing mathematical expressions on the web is Mathematical Markup Language (MathML). It is a format recommended by the W3C Math Working Group as a standard to represent mathematical expressions. MathML is an XML application for describing mathematical notations and encoding mathematical content within a text format. MathML has encoding of two types: content-based encoding, called Content MathML and presentation-based encoding, called Presentation MathML. Content MathML addresses the semantic meaning whereas Presentation MathML addresses the display structure of the mathematical expressions. Throughout this dissertation, the term "semantic enrichment" of mathematical expressions is used to refer to "adding Content MathML markup to Presentation MathML markup of mathematical expressions".

This dissertation presents a method for "semantic enrichment of mathematical expressions" based on Statistical Machine Translation (SMT) approach. SMT is the most widely studied machine translation method. SMT uses a large dataset of good translations, which comprises a corpus of texts already translated into another language called a parallel corpus. It uses these parallel texts to infer a statistical model of translation automatically. The dissertation first formulated the problem as a translation from Presentation MathML to Content MathML, and used MathML parallel markup dataset can be used as a training corpus for SMT. To deal with the structural difference between the two MathML representations, the dissertation proposes a type of rule, named the "segmentation rule". A Support Vector Machines (SVM) classifier is used to disambiguate the sense of mathematical terms. Features for the SVM are extracted from both the presentation of mathematical expressions and their surrounding text.

This dissertation also introduces a SMT-based method that uses a MathML parallel markup corpus to generate pseudo training datasets for the problem of mathematical term

sense disambiguation. Experimental results showed that the proposed method can generate such data automatically and with reasonable accuracy. Based on the dataset generated, a Support Vector Machines classifier is used to disambiguate the sense of mathematical terms. This approach, combined with the statistical machine translation approach, improves the semantic enrichment of mathematical expressions performance.

This dissertation also presents content-based mathematical search system enhanced with the proposed semantic enrichment method. By ascertaining the underlying semantic meanings of mathematical expressions, a mathematical search system is expected to yield better results. Through the experimental study using different types of queries, the dissertation points out the importance to select an appropriate search strategy depending on the search context: Presentation MathML-based search systems are more suitable for elementary functions with less ambiguity in the Presentation MathML expression. Content MathML-based search systems are preferable when dealing with functions with domain-specific definitions. Another situation in which Content MathML works better is when there exist many Presentation MathML representations for a single Content MathML expression.

This research is the first attempt for a quantitative evaluation of a semantic enrichment of mathematical expressions using a manually constructed dataset. The research also proposed two standard metrics for evaluating the task: the tree-edit-distance error rate and the perfect translation rate. Based on the constructed dataset and evaluation metrics, the research firstly showed the potential improvement of mathematical search system using semantic enrichment. The results of this research will serve as a base for future studies in this field.

(Separate Form 3)

博士論文の審査結果の要旨
Summary of the results of the doctoral thesis screening

5 名の審査委員全員出席の下，博士請求論文（以下，「論文」）の内容についてスライドを用いて約 45 分間の口頭発表を行った後，質疑応答を 30 分間行った．

論文では，文書内の数式を検索するための数式表現の意味解析手法について論じている．「数式」は科学技術文書には必須の構成要素であるが，独自の表現形式を持つため，既存の検索手法ではうまく扱うことができない．そこで論文中では数式の意味構造を抽出するための手法を提案し，それに基づく検索性能の改善と課題について述べている．

論文は英語で記述されており，以下の通り全 6 章から構成されている．

第 1 章はイントロダクションである．まず数式検索の必要性や研究動向を概観し，意味的に類似する数式を見つける上での課題を明らかにしている．

第 2 章では，基本概念や関連研究について述べている．特に数式の表現方法として，数式の画面上での表示方法を定義する表示構造表現と，関数と引数など数式構成要素の意味的関係を定義する意味構造表現の 2 つを取り上げ，一般的な表記法である前者を後者に変換する操作を，数式の意味解析として定義している．また，両者の表現形式それぞれについて，これまでに提案された数式検索システムをまとめている．

第 3 章では，数式の表示構造および意味構造を表す 2 つの XML 木構造データを，自然言語処理における機械翻訳手法を用いて対応付ける手法について述べている．複雑な木構造を扱うために，通常の翻訳ルールに加えて，部分木どうしを対応づける分割ルールを新たに導入した上で，あらかじめ両者が対応づけられた訓練用データからこれらのルールを自動獲得する手法を提案して有効性を確認している．

第 4 章では，翻訳の過程で生じる数学記号の意味あいまい性を解消する手法について述べている．意味あいまい性解消ではサポートベクタマシン等の機械学習の適用が有効であることが知られているが，数式のように対象領域が小さくかつ多岐に渡る問題において，正解付きの訓練データを生成することは容易ではない．そこで，第 3 章の機械翻訳手法を適用して擬似的な正解データを自動生成する手法を提案し，訓練データとして人手による正解データを用いる場合と分類性能を比較することで有用性を示している．

第 5 章では，数式検索における数式意味構造の活用について論じている．これまで，共通の数式検索タスクのもとで表示構造表現と意味構造表現を比較する試みはほとんど行われていなかったため，まず両者について検索システムが出力する検索性能を調べ後者の有用性を示している．さらに，論文で提案した意味解析手法の適用により得られる意味構造の有用性についても比較分析を行っている．

第 6 章は結論である．論文の内容を統括するとともに，実用的な数式検索に向けて，今後解決するべき課題や研究の展開を論じている．

口頭発表では論文の内容に沿って，まず研究の背景および位置づけを述べた後に，機械翻訳手法の適用による数式の意味解析，および，数学記号のあいまい性解消手法に関する説明があった．次に，数式検索におけるこれらの手法の有用性について実験に基づく報告があり，最後に，考察および研究全体のまとめを行った．

(Separate Form 3)

質疑応答では，論文及び口頭発表の内容に関して，理論解析に関する定義や証明を中心に質問があり，的確な回答がなされた．論文の内容は電子情報通信学会英文誌に原著論文として採録されており，また一部の内容は International Joint Conference on Natural Language (IJCNLP 2013)などの査読付き国際会議において発表済である．

以上に基づき審査した結果，5 名の審査委員全員一致で，本学位請求論文は学位を授与するのに十分なレベルであるものと判定された．