氏　　名	Néstor ÁLVARO GRADILLAS
学位(専攻分野)	博士(情報学)
学 位 記 番 号	総研大甲第 1886 号
学位授与の日付	平成28年9月28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	Analysis of the Formality of Text and its Impact on Pharmacovigilance Systems
論 文 審 査 委 員	主　査　准教授　宮尾 祐介 准教授　北本 朝展 教授　武田 英明 Principal Research Associate NIGEL COLLIE University of Cambridge 准教授　鶴岡 慶雅 東京大学

(別紙様式 2)
(Separate Form 2)

論文内容の要旨
Summary of thesis contents

This thesis aims to answer the question of whether drug use reports obtained from formal and informal sources have noticeable differences in their formality, and also assess whether these differences in the formality can provide gains to pharmacovigilance systems. Working towards these goals we made three clear contributions that are the development of new resources, i.e. corpora, to perform our studies, a linguistic analysis of the differences between formal and informal pharmacovigilance reports, and an assessment on the impact of the register-related features in pharmacovigilance systems.

The first contribution is motivated after finding that there is no repository meeting our requirements, this is, a data set composed of sentences retrieved from academic texts and from social media messages that contain reports on the use of a closed set of drugs that could be used in our linguistic study to provide our second contribution.

Our second contribution focuses on the linguistic register where we explore it by using a well established method (Multidimensional analysis) from the area of linguistics that is known for being used to evaluate differences in the formality of texts. By using the multidimensional analysis (MD) proposed by Biber we are able to study the register from a more inclusive point of view as this method does not only account for a set of traits, but it also characterizes the texts using different combinations of the studied features so that other elements such as the ``abstractness'' or the ``narrativeness'' of the texts are assessed. The assessment on the differences between generic tweets and drug-related tweets shows that even if both of them belong in the same type of register Biber's schema is able to capture the differences and helps in telling apart the drug use reports due to their higher level of informativeness. Similarly, an analysis comparing the set of drug-related tweets and the corpus of PubMed sentences shows that the academic texts have more traits of ``Information Productions'', proving that the MD analysis can capture those characteristics.

For our third contribution we use a set of features able to capture differences in the register and explore the power of those features in drug safety systems. For that we prepare four different binary classifiers for the tasks of detecting sentences containing first-hand experience reports on the drug use, sentences containing beneficial outcomes, sentences containing negative outcomes and sentences containing any type of outcome (positive or negative) related to the drug use. We also build a named entity recognition (NER) system to detect the mentions of drugs and diseases and symptoms. Those systems also explore additional set of features that MD analysis did not assess, but which are known to carry important formality

information. Those experiments show that for the set of assessed classifiers using PubMed texts the use of Biber features, the use of custom expansion to those features, nor the use of our set of Politeness features can beat the other configurations of the classifiers that do not include such register-related information. We can see, though, that the two of the classifiers for Twitter data do benefit from the use of our custom set of Biber features and from the use of our set of Politeness features as those two sets of features provide significant gains to the baseline system. On the other hand, when evaluating NER systems targeting at the identification of drugs and symptoms and diseases we saw that the set of Politeness features combined with the baseline and word2vec features scored the best result in both PubMed and Twitter data sets, showing that these features can contribute to pharmacovigilance systems.

(別紙様式 3)
(Separate Form 3)

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

博士論文審査は、審査委員出席のもと、7 月 20 日と 7 月 26 日の 2 回に分けて行われた。まず出願者によるプレゼンテーションを 45 分行い、その後質疑応答を 30 分程度行った。本論文は、医薬品安全性監視システムのためのテキストマイニングにおいて、自然言語テキストの形式性に着目した分析とその応用に関する研究である。学術論文データベース PubMed のアブストラクトとソーシャルメディア Twitter のテキストを対象とし、言語学的 register（口語体・文語体など使用状況によって言語表現が変わること）の定量的分析によりこれらのテキストの性質を分析、さらにその知見を応用してテキスト分類や情報抽出の精度を向上すること目指している。

第 1 章では、医薬品安全性監視システムのために様々な自然言語テキストデータから情報収集することが必要であることが主張され、本研究のターゲットとして PubMed と Twitter が挙げられている。そして、本研究では、言語学的 register の定量的分析を行い、これらのテキストの形式性の違いを明らかにすること、そしてその知見を応用してテキスト分類・情報抽出の精度向上を目指すことを提案している。

第 2 章では、自然言語テキストの形式性を定量的に測定する手法として Biber の手法について説明し、これを応用した言語学・自然言語処理の既存研究について概説している。さらに、医薬品安全性監視システムのためのテキストマイニング・自然言語処理技術の既存研究について概説している。

第 3 章では、研究を進めるためのデータとして、特定の医薬品に言及している PubMed アブストラクトおよび tweet を収集し、医薬品、疾患、症状などのアノテーションを付与したデータの開発について報告している。一部のアノテーション（例えば疾患と症状）については区別が困難で作業者一致率が低かったが、これらをまとめる処理を行った結果、作業者間一致率の高い高品質なデータが開発されたことが報告されている。

第 4 章では、このデータと Biber の手法を用いて、一般の tweet、医薬品に言及している tweet、PubMed アブストラクトの言語学的 register の分析を行っている。様々な知見が議論されているが、特に、一般の tweet と医薬品に言及した tweet は多くの点で性質が異なり、後者はよりフォーマルテキストに近いこと、医薬品に言及した tweet と PubMed アブストラクトは多くの点で類似していること、などが示された。

第 5 章では、第 4 章で利用した Biber の手法を応用し、5 種類のテキスト分類・情報抽出タスクにおいて、言語学的 register に関する特徴量を新たに導入する実験を行った。その結果、医薬品利用の実体験が書かれた tweet を分類するタスクなどにおいては、新たに導入した特徴量により精度が大きく向上することが示された。一方、PubMed で同分類を行うタスクや、医薬品の効果の positive/negative を分類するタスクでは、有意な精度向上は見られなかった。以上のことから、言語学的 register は、特定のドメイン・タスクにおいて有用であるとの結論が得られた。

第 6 章では、以上の議論をまとめ、将来課題について議論している。

質疑応答では、本研究の着想にいたった動機、本研究の成果をより多くの医薬品やアプリケーションに応用する可能性、言語学的 register の分析方法の詳細や特定タスクの精度を向上させた原因などについて質問がなされ、おおむね的確な回答が得られた。

審議を行った結果、出願者は情報学分野の十分な知識と研究能力を持つと認められ、また研究内容は学位論文として十分なレベルの新規性や有効性があると認められた。また、

(別紙様式 3)
(Separate Form 3)

本論文の一部の内容について査読付きジャーナル論文 1 編が採録済みである。以上から、審査委員全員一致で、本審査を合格とするとの結論に至った。