語クラスターとランキングモデルを用いる情報更新タスクの扱いに関する研究

Pham Quang Nhat Minh(0810054)
情報科学研究科
北陸先端科学技術大学院大学
2010 年 2 月 9 日

キーワード:情報更新,情報検索,ワードクラスタリンえランキングモデル

文書の頻繁な更新が必要な応用において、情報の更新タスクは重要である。法令ドメインでは、法令の頻繁な更新や相互参照の問題により、情報の更新タスクは重要なものとなる。本論文は,情報更新タスクとして、新しい情報を文書のどこにおけばよいかを決める情報挿入タスクを取り上げ、その方法と実験結果を示す。

従来、情報挿入タスクを階層的ランキング問題として扱った研究がある。この研究では,文書は節やパラグラフの階層として表現され、挿入はその階層木の上で扱われる。新しい情報(文)を文書のどのパラグラフに置くのが最もよいかを決定するために、ランキング関数により全パラグラフを順序付け、最高のスコアのパラグラフが選択される。ランキング関数の値は、挿入文とパラグラフを入力に、訓練データから学習される重み付きベクトルにより計算される。訓練過程は、パーセプトロンアルゴリズムによるオンライン学習として実装されている。

本研究では、二つのデータセットについて、情報挿入タスクのランキングモデルを研究した。従来研究が公開するウィキペディアの挿入データセットおよび本研究で構築した法令データセットである。法令データセットは米国連邦法から作ったものである。

自然言語処理では、二つのテキストセグメントの意味的類似性を測るのに、語と語の意味関係が用いられる。本研究では、二つのテキストセグメントの間の話題に関する重複を測る方法を提案した。具体的には,語のクラスターを導入し、その類似尺度を学習モデルの意味素性として用いる。この方法では、まず、語クラスターを注釈なしのデータから求める。抽出された語クラスターは、表層の形は異なっていても意味的に関連している語の間の意味的類似性および意味的関連性を利用するための中間表現として用いられる。テキストの意味的類似性のスコアは様々な種類の類似性関数により計算される。本研究の結果、語クラスターに基づく素性とベァスラインの素性とを組合せることにより、二つのデータセットを対象にする情報挿入の性能は高くなることが示された。