【公報種別】特許法第１７条の２の規定による補正の掲載
【部門区分】第６部門第３区分
【発行日】平成20年5月15日(2008.5.15)

【公開番号】特開2005-293580(P2005-293580A)
【公開日】平成17年10月20日(2005.10.20)
【年通号数】公開・登録公報2005-041
【出願番号】特願2005-92423(P2005-92423)
【国際特許分類】
Ｇ０６Ｆ　17/27　　(2006.01)
Ｇ１０Ｌ　15/18　　(2006.01)
【ＦＩ】
Ｇ０６Ｆ　17/27　　Ｚ
Ｇ１０Ｌ　 3/00　　５３７Ｄ

【手続補正書】
【提出日】平成20年3月28日(2008.3.28)
【手続補正１】
【補正対象書類名】特許請求の範囲
【補正対象項目名】全文
【補正方法】変更
【補正の内容】
【特許請求の範囲】
【請求項１】
削除補間言語モデルのパラメータを記憶する方法であって、
前記削除補間言語モデル用のパラメータのセットを得るステップと、
前記削除補間言語モデル用の少なくとも１つのパラメータをバックオフ言語モデル用のパラメータとして記憶するステップとを含み、
前記削除補間言語モデルのパラメータにより、Ｎグラム確率を、前記Ｎグラム確率の相対頻度推定値と、より低次のｎグラムの確率との線形補間として決定することが可能であり、
前記バックオフ言語モデルは、前記Ｎグラム確率をより低次のｎグラム、および前記バックオフ言語モデル中で突き止めることができない、任意のＮグラムに対するバックオフ重みで置き換えることを特徴とする方法。
【請求項２】
前記削除補間言語モデル用の少なくとも１つのパラメータを記憶するステップは、単語シーケンスの補間済み確率を前記バックオフ言語モデル中の単語シーケンスの確率として記憶するステップを含むことを特徴とする請求項１に記載の方法。
【請求項３】
前記補間済み確率を記憶するステップは、前記単語シーケンスの相対頻度がしきい値よりも大きいと判定した後で前記補間済み確率を確率として記憶するステップを含むことを特徴とする請求項２に記載の方法。
【請求項４】
前記相対頻度は、小数値を有する頻度カウントに基づいて決定されることを特徴とする請求項３に記載の方法。
【請求項５】
前記補間済み確率を記憶するステップは、前記単語シーケンスが前記バックオフ言語モデル中のｎグラムに対するコンテキストを形成すると判定された後で前記補間済み確率を確率として記憶するステップを含むことを特徴とする請求項２に記載の方法。
【請求項６】

前記削除補間言語モデル用の少なくとも１つのパラメータを記憶するステップは、前記削除補間モデル用の補間重みを前記バックオフ言語モデル用のバックオフ重みとして記憶するステップを含むことを特徴とする請求項１に記載の方法。
【請求項７】
前記補間重みを記憶するステップはさらに、前記補間重みに関連する単語シーケンスを前記補間重みと同じエントリに記憶するステップを含むことを特徴とする請求項６に記載の方法。
【請求項８】
前記パラメータのセットを得るステップは、補間重みのセットを訓練するステップを含むことを特徴とする請求項１に記載の方法。
【請求項９】
前記補間重みのセットを訓練するステップは、頻度カウント範囲のセットごとに別々の重みを訓練するステップを含むことを特徴とする請求項８に記載の方法。
【請求項１０】
前記削除補間言語モデル用の少なくとも１つのパラメータを記憶するステップは、バックオフ言語モデル用のＡＲＰＡフォーマットに準拠するデータ構造を生み出すように前記少なくとも１つのパラメータを記憶するステップを含むことを特徴とする請求項１に記載の方法。
【請求項１１】
コンピュータ実行可能命令を有するコンピュータ可読媒体であって、前記コンピュータ実行可能命令は、
補間の値を通して確率を形成する削除補間言語モデル用のパラメータを識別するステップと、
前記パラメータをバックオフ言語モデル用のバックオフパラメータとしてデータ構造中に配置するステップと
をコンピュータに実行させるための命令であり、
前記バックオフパラメータは、前記Ｎグラムが前記バックオフ言語モデル中で突き止めることができない場合に、重み付けされたより低次のｎグラムの確率をＮグラムの確率に対して代用することを特徴とするコンピュータ可読媒体。
【請求項１２】
前記パラメータをデータ構造中に配置するステップは、前記パラメータが前記バックオフ言語モデルの一部として含まれるべきであると判定するステップを含むことを特徴とする請求項１１に記載のコンピュータ可読媒体。
【請求項１３】
前記パラメータが前記バックオフ言語モデルの一部として含まれるべきであると判定するステップは、訓練テキスト中における単語シーケンスの頻度がしきい値を超えると判定するステップを含むことを特徴とする請求項１２に記載のコンピュータ可読媒体。
【請求項１４】
前記パラメータが前記バックオフ言語モデルの一部として含まれるべきであると判定するステップは、前記パラメータに関連する単語シーケンスが、前記データ構造に記憶されたｎグラム中のコンテキストを形成すると判定するステップを含むことを特徴とする請求項１２に記載のコンピュータ可読媒体。
【請求項１５】
前記パラメータをデータ構造中に配置するステップは、補間済み確率をｎグラムの確率として配置するステップを含むことを特徴とする請求項１１に記載のコンピュータ可読媒体。
【請求項１６】
前記パラメータをデータ構造中に配置するステップは、補間重みをコンテキストに対するバックオフ重みとして配置するステップを含むことを特徴とする請求項１１に記載のコンピュータ可読媒体。

【請求項１７】
前記データ構造はバックオフ言語モデル用のＡＲＰＡ標準に準拠することを特徴とする請求項１１に記載のコンピュータ可読媒体。
【請求項１８】
言語モデルを構築する方法であって、
削除補間を用いて言語モデル用のパラメータを訓練するステップと、
前記訓練されたパラメータの少なくともいくつかを、バックオフ言語モデル用のＡＲＰＡフォーマットに準拠するデータ構造で記憶するステップと
を含むことを特徴とする方法。
【請求項１９】
前記訓練されたパラメータの少なくともいくつかを記憶するステップは、訓練テキスト中でしきい値量よりも多く出現する単語シーケンスに関連するパラメータを記憶するステップを含むことを特徴とする請求項１８に記載の方法。
【請求項２０】
前記訓練されたパラメータの少なくともいくつかを記憶するステップは、前記データ構造に記憶されたｎグラム中のコンテキスト単語として出現する単語シーケンスに関連するパラメータを記憶するステップを含むことを特徴とする請求項１８に記載の方法。