未知の言語にも対応できる統計ベース多言語テキスト処理技術

どんな言語の品詞解析器でも短期間に低コストで構築

市場のグローバル化に伴い，英語以外の文書を解析するニーズが高まっています。東芝は，どんな言語にも短期間に低コストで対応できる，“統計ベース多言語テキスト処理技術”の開発を進めており，今回は，品詞解析器を構築する手法を開発しました。品詞辞書の構築，品詞系列の正解生成，及び文法モデルの学習が技術的なポイントです。この技術を適用したポルトガル語とイタリア語の品詞解析器の品詞推定精度は80～90％でした。

今後は，更に対応言語を増やしていくとともに，より詳細な解析のため，構文解析や固有表現抽出のような高度な言語処理についても取り組んでいく予定です。

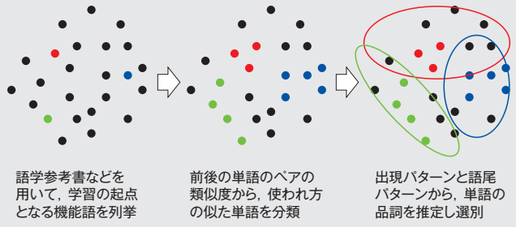

図1．品詞パターンの学習による単語選別 ― 何もメタ情報のない大規模コーパスから品詞ごとのパターンを学習し，それぞれの品詞に相当する単語を選別する。

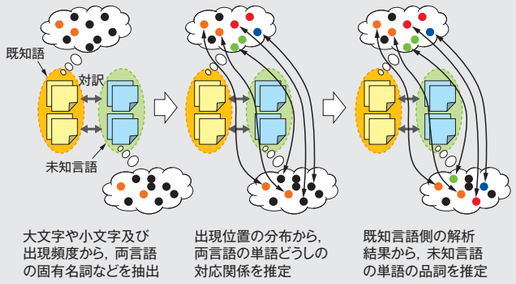

図2．既知言語との対訳による品詞推定 ― 既知言語との対訳コーパスから両言語の単語どうしを対応づけ，既知言語側の情報を元に未知言語側の単語の品詞を推定する。

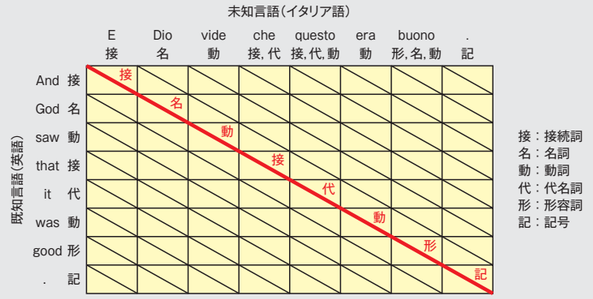

図3．既知言語の品詞系列から推定した未知言語の品詞系列 ― 単語に割り当てられる品詞は一つとは限らないため，対訳の既知言語側における品詞解析の正解から未知言語側の正解を生成し，機械学習の教師データとする。

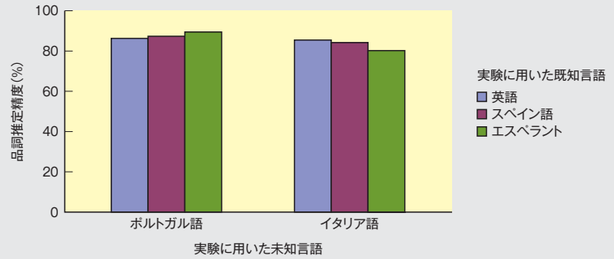

図4．ポルトガル語とイタリア語の品詞解析器の品詞推定精度 ― 英語，スペイン語，及びエスペラントを既知言語として用いて構築された，ポルトガル語とイタリア語の品詞解析器の精度である。半自動で構築したにもかかわらず，元となる既知言語がどれであっても高い精度が得られている。

統計ベース多言語テキスト処理技術とは

近年，市場のグローバル化が進んでおり，世界各地に対応した製品をすばやく提供することが求められています。

東芝はこれまで主に日本語と英語の解析技術に取り組んできましたが，世界各地の言語で書かれた文書が解析できるようになれば，利用者の関心や意図を推定したり，製品に対する評判や苦情を分析したりすることで，様々な製品やサービスを世界各地に展開していくことが可能になります。

しかし，よく知らない言語の解析器を新たに構築しようにも，従来のように言語の特性に合わせて辞書や文法を作り込むのでは言語依存の部分が多く，多大なコストが必要です。

そこで当社は，統計的手法を用いて辞書や文法の構築を半自動化するなど，どんな言語の解析器も低コストですばやく構築できる技術として，“統計ベース多言語テキスト処理技術”の開発を進めています。将来的には，高度な言語処理が行える解析器をも表層情報や統計情報の組合せだけから構築することを目指していますが，まず，第1段階として，主にヨーロッパ系言語の品詞解析器を構築する技術を開発しました。

この結果，よく知らない言語に対しても，キーワード抽出や文書分類など様々な応用がすばやく実現できるようになります。

品詞辞書の構築

この技術の一つ目のポイントは，品詞辞書の構築です。コーパス（言語データ）に出現する単語の名詞らしさや動詞らしさを判定することで品詞を推定します。用いるコーパスの質や量によって適した手法が異なるため，2種類の手法を併用します。

第1の手法（**図1**）では，よく知らない言語（以下，未知言語と呼ぶ）の大規模コーパスから品詞ごとのパターンを学習し，単語を選別します。冠詞や前置詞など，機能語の主なものは学習の起点として人手で与えておきます。

ある単語に対し，前後に出現する単語のペアを出現パターンとすると，似た使われ方をする単語どうしでは，出現パターンは似たようなものになります。そのため，使われ方が似た単語を出現パターンの類似度によって分類できます。しかし，これだけでは精度が粗いため，別途抽出した品詞ごとの語尾パターンも併用して単語を選別します。

第2の手法（**図2**）では，当社が解析器を持っている言語（以下，既知言語と呼ぶ）との対訳コーパスから両言語の単語どうしを対応づけ，未知言語の単語の品詞を推定します。

まず，文字種や出現頻度などに基づき両言語の固有名詞を抽出します。また，未知言語側から語尾が異なる単語の集合を語形変化候補として抽出します。

次に，対訳での出現位置などに基づき両言語の単語どうしの対応を推定します。ヨーロッパ系言語どうしは，日本語と比べると語順がかなり似ており，対応する単語どうしの出現位置は似たようなものになるためです。

その他，既知言語の標準形と未知言語の語形変化候補から未知言語の語形変化を求め，品詞を推定します。

文法モデルの機械学習

この技術の二つ目のポイントは，文法モデルの機械学習です。

それぞれの単語に割り当てられる品詞は一つとは限らないため，品詞解析結果の候補は一般に複数あります。それらの中からもっとも確からしい候補を選択するためには文法モデルが必要ですが，未知言語に対して人手で記述することは不可能です。また，機械学習のための教師データとなる品詞解析の正解を，未知言語に対して人手で多数与えることは困難です。

一方，ヨーロッパ系言語は，基本的な文法に大きな差がないとされています。そこで当社は，品詞や単語の対応ごとにコストを設定した最短経路問題を解くことによって，対訳の既知言語側を品詞解析して得られた正解から未知言語側の正解を生成し（**図3**），それらに基づいて文法モデルを機械学習することにしました。ただし，両言語で語順が大きく入れ替わっていると正解をうまく生成できないため，更に改善していく必要があります。

ポルトガル語とイタリア語への適用

この技術を適用して構築した，ポルトガル語とイタリア語の品詞解析器の品詞推定精度を**図4**に示します。大規模コーパスとしてWikipediaの記事全文を，また，対訳コーパスとして聖書を用いて品詞辞書を構築し，当社が解析器を持っている英語，スペイン語，及びエスペラントの情報を用いて文法モデルを機械学習したものです。

半自動で構築したにもかかわらず，80～90％の高い精度が得られています。また，元となる既知言語がどれであってもほぼ同じ精度が出ており，言語の特性に依存しない安定した手法と言えます。

今後の展望

今回はポルトガル語とイタリア語を対象に品詞解析器を構築しましたが，今後はロシア，ヒンディー，アラビア語など新興国の言語に拡張していきます。また，品詞解析にとどまらず，より高度な言語処理についても技術開発を進めていきます。

文　献

(1) 山崎智弘 他．“統計的手法に基づく品詞解析器の半自動構築”．第9回日本データベース学会年次大会．静岡県，2011-02，日本データベース学会 他．A8-1.

山崎　智弘
研究開発センター
知識メディアラボラトリー研究主務