コーパスに基づく機械翻訳に関する研究

土居, 誉生

(Degree)
博士（工学）

(Date of Degree)
2007-03-25

(Date of Publication)
2012-04-09

(Resource Type)
doctoral thesis

(Report Number)
甲3965

(URL)
https://hdl.handle.net/20.500.14094/D1003965

【 358 】

氏　　名・（本　籍）	土居　誉生	（　高知県　）
博士の専攻分野の名称	博士（工学）	
学　位　記　番　号	博い第449号	
学位授与の　要　件	学位規則第５条第１項該当	
学位授与の　日　付	平成19年3月25日	

【 学位論文題目 】

コーパスに基づく機械翻訳に関する研究

審　査　委　員

主 査　教　授　隅田　英一郎
教　授　上原　邦昭
教　授　田村　直之

近年，対訳コーパスから自動的に翻訳システムを構築するコーパスベースの機械翻訳技術の研究開発が盛んになってきており，実用化に向けた一層の性能向上が期待されている．一般に機械翻訳システムの性能向上のためには，翻訳エンジン，翻訳前処理，翻訳後処理の 3 側面での工夫による効果が期待できる．本論文では，グラフベースの用例検索手法を使った翻訳エンジンの話題を中心に，コーパスベース翻訳エンジン共通の課題への前・後処理による対策を取り上げ，コーパスベース翻訳の性能向上について論じる．

用例翻訳や統計翻訳といったコーパスベース翻訳では，基本的に単語や句などの小さな単位の翻訳知識を獲得しそれを使って翻訳を実行する．一方，翻訳すべき入力文とほとんど同じ文とその訳文のペアがコーパス中に存在すれば，そのペアを直接利用することで非常に良い訳が得られる．この原理に基づく文単位の用例翻訳を想定し，その実現に必要な課題に取り組む．用例翻訳では，入力表現に最も類似した対訳用例を抽出し，その翻訳表現の一部を変更して入力に対する訳を生成する．通常，対訳表現の単位は句である．句の組合せにより多くの入力文に対応可能であるが，各部分の誤りの合成や組合せの不整合により，不適切で不自然な訳文が生成されてしまう危険がある．一方，用例単位として文を使う場合，その危険性を抑えることができる．入力文全体に類似した文単位の用例が見つかれば，正確で自然な翻訳文を得ることができる．もちろん文は句に比べ，より長い単位であり汎用性が劣るため，文単位の用例翻訳はカバレッジの面で短所を持つ．文単位の用例を使って十分な翻訳カバレッジを得るためには，大規模な対訳コーパスを用意しなければならず，大規模コーパスから用例を検索するための効率的な検索手法が必須となる(第2章)．

ここでは編集距離に基づく類似度を用いた文単位の用例翻訳システムを想定し，そのシステムが大規模コーパスを扱うための効率的な検索手法を提案する．対象となる検索処理は，コーパス中の原言語文を候補文とし，入力文との距離が閾値以内で最小の候補文を全て求めることである．その目的のために全ての候補文について逐次的に距離を計算するのでは時間がかかり過ぎてしまう．そこで候補文集合の分割，単語グラフ，A*アルゴリズムを利用した効率的な検索手法を提案する．提案手法では 2 文間の逐次的な照合は行わず，グラフ化された複数の候補文と入力文との照合を同時並行的に進める．この手法では編集距離の定義と与えられた閾値に関して検索もれはない．提案検索手法では，内容語数と機能語数を基に候補文をグループ分けする．これにより入力文の内容語数と機能語数および距離閾値を基にしたグループ単位での枝刈りが可能となる．各グループ毎に複数の候補文が一つの単語グラフにまとめられる．単語グラフは有向グラフであり，先頭ノードから最終ノードに至る可能な道筋と候補文が互いに対応する．複数の文に共通な単語列がグラフ中で一つにまとめられ，ノード数が最小となるように圧縮される．グループ内の検索は，単語グラフの先頭ノードから最終ノードまでの可能な全経路について，各経路に現れる単語列と入力単語列との照合結果の中から編集距離を最小にするものを探索することである．この探索問題の解法に A*アルゴリズムを用いる．一般に A*アルゴリズムでは，問題状態集合の中から最終コストの下限の推定値が最小のものが選ばれ継続状態に展開される．ここで対象とする問題では，状態は，単語グラフの経路と入力文との照合の途中経過を意味する．コスト推定の際，単語グラフ内では全ての候補文の単語数が等しいという条件が利用される(第 3 章)．

旅行会話に関する数十万文規模の対訳コーパスを使った日英翻訳実験を通して，提案検索手法でもって実装した翻訳システムの性能を評価した．利用したコーパスは，海外旅行者向けのフレーズブックに相当する内容の日本語文とその英訳からなる．翻訳結果の品質評価のために客観評価スコアと主観評価ランクを用いた．また処理効率の評価のために，Pentium4/2GHz の通常のパーソナル・コンピュータ上での翻訳処理時間を計測した．15 万対訳からなる学習セットを使って基本的な翻訳性能を評価した結果，最高ランクに分類される訳文の割合が 71%と大きく，高い翻訳品質が示された．翻訳処理時間は，平均 0.2 秒，最大 3.3 秒であり，提案した用例検索手法により効率的な処理が実現された．コーパスサイズと翻訳性能の関係を調べるために，学習セットの大きさを 1.9 万から 30 万対訳の間で変えて性能評価を行った．結果として各指標は，コーパスサイズが大きくなるほど大きなカバレッジと高い翻訳品質が得られることを示した．一方，処理時間はコーパスサイズのほぼ 1/2 乗のオーダーに抑えられている．編集距離基準を用いた文単位の用例翻訳システムは，大規模コーパスを使うことにより高品質の翻訳能力を持ち，提案検索手法を使うことにより効率的な翻訳処理が可能となっている(第 4 章)．

一般に機械翻訳システムは入力文が長くなると誤りが多くなる．第 4 章の実験でもその傾向が確認されている．しかし入力文が長くて翻訳誤りが起こる場合でも，分割した各部分に対しては翻訳が成功する可能性がある．入力文を互いに独立性の高い部分に分割できれば，それぞれの翻訳結果を同じ順番に並べることにより入力全体を翻訳することができる．そこで，機械翻訳システムを使ったより良い翻訳が可能となるように，前処理で入力文を分割することにより翻訳システムに渡す単位を適正化するアプローチをとる．文分割に関する従来の研究では，分割点周辺の N グラムなどの単語の連接の特徴に基づいた手法が多い．それに対して提案手法では，単語連接に基づく分割を補うための別指標として類似度を導入する．当分割手法では，N グラムに基づき分割点の候補を生成し，コーパスを使って計算した類似度に基づき分割点の最良の組合せを選択する．この選択は，コーパスベース翻訳システムは，学習コーパスに存在する文と類似した文は正しく翻訳することができるという仮定に基づいている．当手法の評価のため，句単位の用例翻訳，文単位の用例翻訳，統計翻訳の異なる手法による 3 つのコーパスベース翻訳システムによる英日翻訳実験

を行った．実験結果は，いずれのシステムに対しても提案分割手法が有効であり，類似度の使用は翻訳品質を改善することを示している(第５章)．

コーパスベース翻訳では，特徴的な翻訳誤りとして，入力文と関連のない語が訳文に湧き出す問題が観察される．第 4 章の実験結果からも翻訳誤りの多くがこのタイプの誤りに分類される．この問題は単語アラインメントというコーパスベース翻訳に本質的な処理に由来する．この湧き出し誤りへの対策として，後編集により自動修正するアプローチを提案する．後編集アプローチには，特定の翻訳システムに限らず適用可能という利点がある．提案手法は誤り語を自動で削除する．この手法では，まず異なる言語の単語間の対応確率を示す単語翻訳モデルを利用し，出現の期待値が閾値以下の訳語を誤り語候補として検出する．次に対訳コーパスから得られた用例を利用して誤り語候補の正しさを検証する．つまり誤り語候補が特定の類似用例に存在し，その用例において入力文と共通する語句から相対的に高い期待値が得られれば，その誤り語候補は誤りではなく正しい訳語と判断される．最終的に残った誤り語が訳文から削除される．日英および中英翻訳を対象とした複数のシステムを使った実験で，誤り語自動削除による翻訳精度向上効果が確認された(第６章)．

以上のように，本論文ではコーパスベース翻訳の性能向上のための手法について論じた．単語グラフを使った用例検索手法により高品質で高速な翻訳システムを実現することができた．さらに関連する翻訳誤りに対して，前処理による入力文分割，後処理による湧き出し語削除の対応策を提案し，複数の翻訳システムにおいて効果を確認した．

593

（別紙１）　　論文審査の結果の要旨

氏名	土居　誉生		
論文題目	コーパスに基づく機械翻訳に関する研究		
審査委員	区　分	職　名	氏　　名
	主　査	教授	隅田　英一郎
	副　査	教授	上原　邦昭
	副　査	教授	田村　直之
	副　査		
	副　査		

要　　旨

本論文は、現在の機械翻訳の中心技術であるコーパスベース翻訳手法（対訳コーパスから自動的に知識を獲得し、これに基づいて翻訳する手法）に関連して、翻訳処理、同前処理、同後処理の３側面において、新らたなアルゴリズムを提案し、その有効性を実証的に示した。

論文の構成は以下の通りである。

第２、３、４章において、代表的コーパスベース翻訳手法である用例翻訳に関して、特に、用例を文単位で検索して利用する手法の実現に必要な課題に取り組んでいる。文単位の用例翻訳は、入力文全体に類似した文単位の用例が見つかれば、正確で自然な翻訳文を得ることができる。一方、文単位の用例は、より短い単位に比べて他の条件が同じ場合に検索ヒット率が相対的に低いという短所を持つ。したがって、文単位で入力に対する十分なカバレッジを得るためには、大規模な対訳コーパスを用意する必要がある。大規模なコーパスを使った場合に、効率的に用例を検索する手法が不可欠な課題となる。

提案手法では、対訳コーパス中の原言語文を候補文とし、入力文との編集距離が最小の候補文を全て求める。そのために全ての候補文について逐次的に距離を計算するのでは実用的な反応時間は実現できない。そこで候補文集合の分割，単語グラフ、A*アルゴリズムを利用した効率的な検索手法を提案した。提案手法では２文間の逐次的な照合は行わず、グラフ化された複数の候補文と入力文との照合を同時並行的に進める（第３章)。

旅行会話に関する対訳コーパスを使って評価している。15 万対訳からなる学習セットにおいて、良質の訳文の割合が 71%と大きく、高い翻訳品質が示された。翻訳処理時間は、Pentium4/2GHz の通常のパーソナル・コンピュータで、平均０．２秒、最大３．３秒であり、提案した用例検索手法により効率的な処理が実現されている。コーパスサイズとの関係を調べると、コーパスサイズが大きくなるほど大きなカバレッジと良い翻訳品質が得られることが確認できた。一方、コーパスサイズを大きした場合の処理時間は、コーパスサイズのほぼ 1/2 乗のオーダーに抑えられている。提案検索手法を使うことにより、高品質で、また効率的な翻訳処理が可能となった(第４章)。

第５章では、前処理に取り組んでいる。一般に機械翻訳システムは入力文が長くなると誤りが多くなる。しかし、入力文を互いに独立性の高い部分に分割できれば、それぞれの翻訳結果を同じ順番に並べることにより入力全体を翻訳できる可能性がある。文分割に関する従来の研究では、分割点周辺の N グラムなどの単語の連接の特徴に基づいた手法がある。この単語連接に基づく分割を補うための別指標として類似度を導入する。当分割手法では、N グラムに基づき分割点の候補を生成し、コーパスを使って計算した類似度に基づき分割点の最良の組合せを選択する。

このコーパスベースの前処理手法の評価のため、句単位用例翻訳、文単位用例翻訳、統計翻訳の異なる手法に基づく３つのコーパスベース機械翻訳システムを対象として英日翻訳実験を行った。実験結果は、どのシステムに対しても提案分割手法が有効であり、類似度の使用は翻訳品質を改善することを示している。

第６章では、後処理に取り組んでいる。コーパスベース翻訳の特徴的な翻訳誤りとして、入力文と関連のない語が訳文に湧き出す問題が観察される。この問題は自動単語アラインメントというコーパスベース翻訳に本質的な処理における失敗に由来する。この湧き出し誤りへの対処として、後編集により自動修正アプローチを提案している。

氏名	土居　誉生

この手法では、まず異なる言語の単語間の対応確率を示す単語翻訳モデルを利用し、出現の期待値が閾値以下の訳語を誤り語候補として検出する。次に対訳コーパスから得られた用例を利用して誤り語候補の正しさを検証する。つまり誤り語候補が特定の類似用例に存在し、その用例において入力文と共通する語句から相対的に高い期待値が得られれば、その誤り語候補は誤りではなく正しい訳語と判断される。最終的に残った誤り語が訳文から削除される。

IWSLT04 という公開の機械翻訳評価型ワークショップに参加した複数のシステムを使って評価実験を行った。日英および中英翻訳を対象としたシステムを使った実験で、提案の誤り語自動削除による翻訳精度向上効果が確認された。

本論文ではコーパスベース翻訳の性能向上のための手法について論じている。

大規模コーパスの利用を可能とする効率的な用例検索手法により高品質で高速な翻訳システムを実現することができた。さらに関連する翻訳誤りに対して、前処理による入力文分割、後処理による誤り語削除の対応策を提案し、複数の翻訳システムにおいて効果を確認した。以上のように、提案された手法は全て実装され、大規模な言語データを用いて、機械翻訳の翻訳品質を向上するために有効であることが確認されている。

本論文では、コーパスベース機械翻訳に対して、様々な視点からの翻訳品質改善を可能とする技術を提案しており、機械翻訳システムの構築に関して重要な知見を得たものとして価値ある集積だと認める。

よって，学位申請者の　土居　誉生　は，博士（工学）の学位を得る資格あると認める．