様式 C-19

科学研究費補助金研究成果報告書

平成２１年　５月２０日現在

研究種目：基盤研究(C)
研究期間：2007～2008
課題番号：19500114
研究課題名（和文）ベイズ統計学を利用した構文情報に基づく統計的機械翻訳モデルの開発

研究課題名（英文）Development of statistical machine translation model using Bayesian statistics

研究代表者
山本 幹雄（ YAMAMOTO MIKIO ）
筑波大学・大学院システム情報工学研究科・教授
研究者番号：40210562

研究成果の概要：
構文情報を利用した統計的機械翻訳法として代表的な階層フレーズモデルに着目し、日英間のように比較的遠い言語間の翻訳精度を上げる２つの手法を開発した。１つ目は階層フレーズからフレーズ順序知識のみを分離し、より柔軟にフレーズ移動を行えるモデルを開発した。２つ目は単語あるいは翻訳の大きな移動をサポートする特別なルールを含む場合でも効率的にモデル推定を行える階層フレーズモデルのベイズ的推定手法とその利用法を検討／開発した。

交付額

（金額単位：円）

	直接経費	間接経費	合　計
２００７年度	2,200,000	660,000	2,860,000
２００８年度	1,300,000	390,000	1,690,000
年度			
年度			
年度			
総　計	3,500,000	1,050,000	4,550,000

研究分野：　計算機工学
科研費の分科・細目：　情報学・知能情報学
キーワード：　自然言語処理

１．研究開始当初の背景

最近のweb技術の発展により個人的なレベルでの情報の発信と利用がきわめて容易になっているが、いまだ言語の違いは情報流通の大きな壁となっている。特にこれまでは英語を理解できればインターネット上の大部分の情報を利用できたが、今後、開発途上国の大規模なIT化によって様々な言語で大量の情報が供給される可能性が高く、英語と日本語だけでは全世界の情報の多くを取りこぼす可能性が高い。しかし、２ヶ国語以上の外国語を修得することは一般に困難であり、また、多様な言語で供給される情報量の増加に対して比較的マイナーな言語の専門家を十分に育てるのも簡単ではない。このような状況では言語間の翻訳を自動的に行う機械翻訳の技術が今後ますます重要となることは明らかである。

また、ここ１０年ほどで統計的な機械翻訳システムが目覚しく発展してきたが、日本語

-英語間など遠い言語間の翻訳は性能的に不十分である。今後は英語とフランス語などのように比較的近い言語間の翻訳だけではなく、日本語と英語などのように比較的遠い言語間の翻訳性能を上げる必要がある。

２．研究の目的

統計的機械翻訳は翻訳単位を単語からフレーズにすることで近年急速に性能向上してきた。しかし、フレーズの翻訳結果を並び替える部分のモデルは不十分であり、特に日本語-英語間のように並びが大きく異なる言語間の翻訳では性能が顕著に低下することが問題であった。本研究では、構文情報を利用した統計的機械翻訳法として代表的な階層フレーズモデルに着目し、単語あるいはフレーズを翻訳した部分の順序を大きく変更する必要がある言語間の翻訳精度を改良することを目的とする。具体的には、フレーズの長距離移動を柔軟に可能とする手法と長距離移動を許すモデルに対して推定精度の高いモデルを推定するベイズ統計学を利用した手法を開発することを目的とした。

３．研究の方法

１つ目の目的であるフレーズ並び替えを柔軟に行うモデルの開発としては、いくつかのものを試したが、最終的に、原言語側の単語を適用制約として、階層フレーズをフレーズ順序テンプレートとみなす手法を提案・開発した。これによって、任意の２つの原言語フレーズの並びを、目的言語側で順序的にそのままかあるいは交換するかを判断できる場合が増加する。２つ目の目的である頑健なモデル推定手法としては、日本語-英語間のように単語の移動が比較的大きな言語ペアの翻訳において必要となる特別なルールを含む場合でも効率的に推定する階層フレーズモデルのベイズ的推定手法を開発した。以下、それぞれの手法について述べる。

(1) 階層フレーズからのフレーズ並び替えテンプレートの抽出

Chiang(2007)が提案した階層フレーズモデルはフレーズベースの統計的機械翻訳手法に構文的翻訳知識をフレーズの並び替えのために組み込んだモデルであり、今後の発展が期待されている。しかし、Chiangの方法は、フレーズの翻訳とフレーズの並び替えを一つのルールで同時にモデル化している点で、ルールがヒットすればかなり正確な翻訳が期待できるが、ルールの適用条件が厳しいためにヒットしない場合が多いという適用可能性の点で問題があることが指摘されている。

本研究では、階層フレーズの適用可能性が低い点を改善するために、階層フレーズルールより、フレーズの並び替え知識だけを分離／適用するHPART(Hierarchical Phrases As Reordering Templates)と呼ぶ方法を提案／評価した。日本語の「に関する」を英語に翻訳する場合を例に取り、以下に基本的なアイデアを説明する。

「に関する」を「of」に翻訳する際の典型的な階層フレーズルールは以下のようになる。

X →< X_0 に関する X_1, X_1 of X_0 >

このルールは、鍵括弧の中のカンマで区切られた左側が日本語のパターン、右側が英語のパターンを表現しており、日本語における「関する」の前のフレーズX_0が、英語側においては「of」の後ろに出現することを意味している。すなわち、このルールは日本語の「に関する」を「of」に翻訳する際に前後のフレーズの順序が入れ替わることを意味する。これは単語あるいはフレーズの翻訳と同時にフレーズの位置をモデル化しており、強力な翻訳知識であるといえる。しかし、「に関する」に対応する英語単語／フレーズは様々なものが考えられる。例えば、「for」「on」「,」などであり、例を図１に示す。

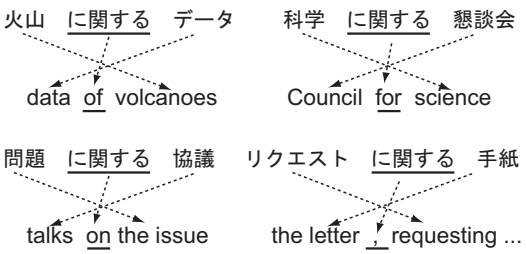

図１： 「に関する」の翻訳例

図１を観察すると、日本語の「に関する」は一般的に様々な単語に翻訳されうるが、前後のフレーズの順序が逆転するという点は一般的であるように見える。しかし、階層フレーズモデルでは、この一般性をモデル化することはできず、それぞれの翻訳単語毎に以下のような階層フレーズルールが学習されている必要がある。

X →< X_0 に関する X_1, X_1 of X_0 >

X →< X_0 に関する X_1, X_1 for X_0 >
X →< X_0 に関する X_1, X_1 on X_0 >
X →< X_0 に関する X_1, X_1 , X_0 >
...

これらの規則は学習データにすべて例が出現していない限り学習はできないため、あらゆる単語／フレーズ毎のルールを学習することは困難である。

本研究で提案するHPARTRは、図１の状況で、「に関する」の前後のフレーズは英語では順序が逆転するという知識をモデル化する。基本的なアイデアは非終端記号を２つ含む階層フレーズルールは必ず非終端記号に対応するフレーズの並び替えの知識を持っているため、その順序情報だけを使えるようにすることである。これは、次のように実現する。再度以下のルールを考える。

X →< X_0 に関する X_1, X_1 of X_0 >

フレーズの順序だけに着目すると、このルールは２つのフレーズが英語側で逆順になることを示している。ただし、制限なしに逆順にはできなので、HPARTでは原言語側に現れる単語を制約とする。すなわち、上記のルールであれば、「に関する」をどちらかに含む２つのフレーズは逆順に翻訳すればよいという知識を表していると考える。このようにすれば、別の知識として「に関して」が「on」「for」「,」などに翻訳できるという知識さえあれば、組合わせ的に柔軟に正しく翻訳できることになる。

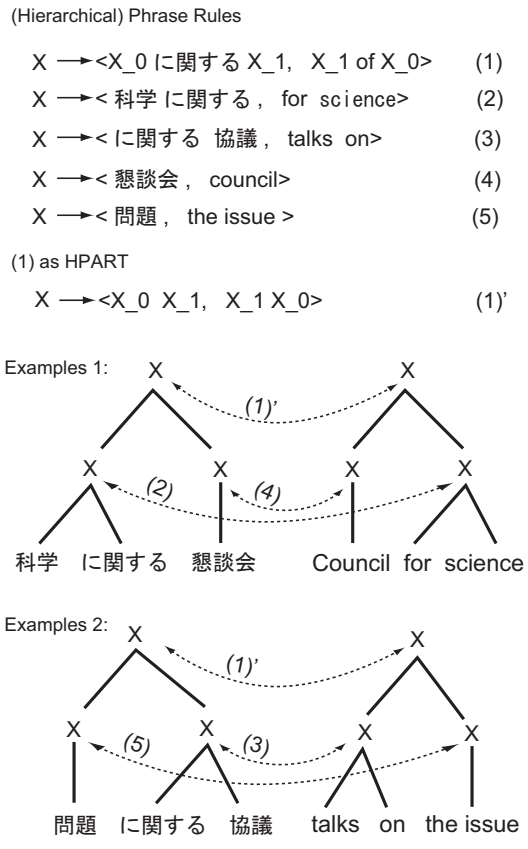

図２： HPARTの動作例

HPARTの動作例を図２に具体的に示す。この中で、(1)の階層フレーズルールから(1)'のHPARTが抽出され、(1)'とその他の翻訳ルール(2)～(5)が組み合わさって、単純な階層フレーズルールでは達成できなかったような柔軟さで２つの例(Example1と2)が正しく翻訳できるようになる。

(2) 階層フレーズモデルの推定手法

階層フレーズモデルの推定にはヒューリスティックスが多用されていたが、2008年、Blunsom ら(2008)によって、EM アルゴリズムとベイズ的な手法を用いて事後確率最大化の原理で確率の推定が効率良く解けることが示された。本研究ではこのBlunsom らの手法に着目し、翻訳モデルの性能改善を目指した。

階層フレーズモデルは語順移動をよくモデル化しており、非常に優れたモデルであるが、Blunsom らの実験においては、中距離の語順移動については間接的にしかモデル化されていなかった。本研究では、この欠点を補うために、フレーズを並べる規則(以下 glue rule と呼ぶ)を導入し、Blunsom の手法をこの glue rule を含んだモデルでも推定可能とするように改良した。

本研究で導入した glue rule は、以下の２つのルールである。

X →< X_0 X_1, X_0 X_1 >
X →< X_0 X_1, X_1 X_0 >

上は原言語文のフレーズ並びを目的言語w側でも保存する接続ルールであり、monotone glue rule と呼ばれる。下のルールは原言語文のフレーズ並びを目的言語側で逆転させて接続するルールであり、inverse glue rule と呼ばれる。特に inverse glue rule を従来のルール集合に加えることで日本語-英語間の翻訳では語順移動を明示的にモデル化できる。

上記の glue rule を加えたモデルに Blunsom らによる手法を適用する場合、glue rule の確率推定のために glue rule が適用可能なすべての構文木を考慮するメカニズムを組み込んだ手法を開発した。しかし、残念ながらこれには推定中に考慮すべき構文木の数が大量に増えることになりかなりの計算時間が必要となる。より具

体的には、図３に示すように、隣接している非終端記号の境界線の位置で対象としている単語列の単語数-1 パターンの場合分けが必要となる点である。図３の例では、glue rule を１回適応するために 63 パターンもの場合分けが必要となっていることがわかる。このような場合分けが、2 単語以上のすべての部分単語列を評価する際に現われるため、効率的な計算のために部分構造をキャッシュし、重複した計算を減らす手法を開発した。glue rule を含めない場合でも、目的言語側のルールに非終端記号が隣接する場合や、対象となる単語列に対して複数パターンの適応法が考えられる場合があった。しかしこのようなケースはむしろ稀なケースであり、2 単語以上の全ての単語列に対して必ず必要となる今回のケースとは全く異るものとして理解されるべきである。この違いが現実的には非常に大きな差となって現われる。

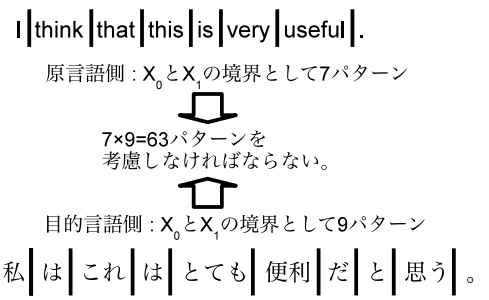

図３：Glue rule を１回適応するのに必要な場合分け

４．研究成果

統計的機械翻訳の性能向上のための２つの手法を提案した。その性能を評価するために、実際に翻訳実験を行い翻訳性能によって手法の評価を行った。また、その他、本研究の成果によって新たに開発した統計的機械翻訳システムを用いて、競争型コンテストの一つである国立情報学研究所主催の NTCIR-7 特許翻訳タスク(Fujii et al. 2008)に参加し、共通的／協調的評価基盤の上での性能評価を行った。様々な制約から、フレーズベースの最高性能のシステムには及ばなかったものの、共通的な評価データを用いた実験によって提案手法の有効性を示すことができた。以下、２つの提案手法の性能評価をまとめる。

(1) 階層フレーズからのフレーズ並び替えテンプレートの抽出の評価

本手法の評価は主に上記で述べた NTCIR-7特許翻訳タスクで提供されたデータを用いて行った。訓練データは、日米の特許データのうち、日本語から英語に翻訳出願されている約10万件から抽出した約180万文の翻訳文ペアからなっている。このデータからフレーズルールを約7000万、階層フレーズを約3000万を自動学習した。テストデータとしては、同タスクで提供された1381文の日本語に対して英語へ翻訳し、BLEUと呼ばれる客観評価指標による評価を行った。

表１が評価結果である。BLEU値は0%～100%の値を取り、大きいほどよい翻訳であることを意味する。「span」は階層フレーズの最大長の制限である。HPと書いあるのがベースラインであり、HP+HPARTと書いてあるものが本研究で提案した手法である。表１より、従来法に比べて1%以上の性能向上が確認できたことが分かる。

表１　HPARTの性能評価結果(BLEU値)

	HP	HP+HPART
span=10	22.84%	24.04%
span=15	23.00%	23.99%

(2) 階層フレーズモデルの推定手法の評価と効果的な利用法

標準的な階層フレーズモデルに語順移動をあまり支援しない monotone glue rule のみを用いた場合の性能と monotone glue rule だけでなく語順移動を積極的に支援する inverse glue rule も加えた場合の性能を比較したところ、性能の向上が見られた。

次に、glue rule を含めたパラメータ推定の結果と glue rule を含めないパラメータ推定の結果を用いて検討を行なった。この結果、Blunsom らの提案した glue rule 無しの推定結果でも概ね前述の glue rule を単純に加えた従来手法よりも良い性能を示した。

さらに、glue rule を加えた推定結果と glue rule 無しの推定結果による翻訳精度を比較した。語順移動を支援しない書換え規則のみの場合、Blunsom らの結果とほぼ同様の結果を示したが、語順移動を積極的に支援する書換え規則を推定の段階から導入する手法においては、glue rule を含まない推定結果よりも有意に高い性能を示

した。

これらの実験結果より、日本語-英語間の翻訳においては中距離の語順移動が性能向上のために重要であることが明らかとなり、モデル推定の段階から中距離の語順移動を考慮することの重要性を示すことができた。

参考文献

Blunsom, P., T.Cohn and M. Osborne. 2008. A discriminative latent variable model for statistical machine translation. In Proc. of ACL-08, pp.200-208.

Chiang, D. 2007. Hierarchical phrase-based translation. Computational Linguistics, 33(2), pp.201-228.

Fujii, A., M. Utiyama, M. Yamamoto, and T. Utsuro. 2008. Overview of the patent translation task at the NTCIR-7. In Proc. of NTCIR-7, pp.389-400.

５．主な発表論文等

（研究代表者、研究分担者及び連携研究者には下線）

〔雑誌論文〕（計２件）

① A.Fujii, M.Utiyama, M.Yamamoto and T.Utsuro, Producing a test collection for patent machine translation in the seventh NTCIR workshop, Proc. of The 6th International Conference on Language Resources and Evaluation, pp.671-674, Morocco, 2008.(査読あり)

② A.Fujii, M.Utiyama, M.Yamamoto and T.Utsuro, Toward the Evaluation of Machine Translation Using Patent Information. The 8th Conference of the Association for Machine Translation in the Americas, pp.97-106, 2008.(査読あり)

〔学会発表〕（計５件）

① 藤井敦、内山将夫、山本幹雄、宇津呂武仁:「大規模な共通基盤による機械翻訳システムの比較評価: NTCIR 特許翻訳タスク最新事情」、言語処理学会第 15 回年次大会発表論文集, pp. 204-207. 2009. 3. 3. 鳥取.

② 越川満、内山将夫、梅谷俊治、松井知己、山本幹雄:「統計的機械翻訳におけるフレーズ対応最適化を用いた翻訳候補のリランキング」、言語処理学会第 15 回年次大会発表論文集, pp. 204-207. 2009. 3. 5. 鳥取.

③ M.Yamamoto et al., Integer programming for a phrase alignment problem on statistical machine translation. Mathematical Programming in the 21st Century：Optimizatioin Modeling and Algorithms, pp.87-93, 2008.7.24. Kyoto.

④ A.Fujii, M.Utiyama, M.Yamamoto and T.Utsuro. Overview of the Patent Translation Task at the NTCIR-7 Workshop. Proceedings of NTCIR-7 Workshop Meeting , pp.389-400. 2008.12.18. Tokyo.

⑤ M.Yamamoto et al., Diversion of hierarchical phrases as reordering templates. Proceedings of NTCIR-7 Workshop Meeting, pp.466-470. 2008. 12.18. Tokyo.

６．研究組織

(1)研究代表者

山本　幹雄（YAMAMOTO MIKIO）

筑波大学・大学院システム情報工学研究科・教授

研究者番号：40210562

(2)研究協力者

乗松　潤矢（NORIMATSU JYUNYA）

筑波大学・大学院システム情報工学研究科・コンピュータサイエンス専攻博士前期課程・学生