様式 C-19

科学研究費補助金研究成果報告書

平成 23 年　3 月 31 日現在

機関番号：34419
研究種目：基盤研究(C)
研究期間：2008 年度～2010 年度
課題番号：20500149
研究課題名（和文）　半教師あり学習による対訳コーパスのアラインメント

研究課題名（英文）　Semi supervised word alignment model for parallel corpus

研究代表者
山本　博史　　（YAMAMOTO HIROIFUMI）
近畿大学・理工学部・教授
研究者番号：00395013

研究成果の概要（和文）：本研究では、対訳文間単語アライメントを決定する際に、従来の単語情報だけでなく、品詞や、構文構造と取り入れることで、精度の向上を図る。この時、全ての学習コーパスに対して品詞や、構文情報を与えることは困難であるため、半教師ありの学習を行う。日英中の 3 ヶ国語パラレルコーパスに対し、品詞タグおよび固有名詞意味情報を用いた半教師ありアライメント学習を行い、効果が確認できた。さらに構文構造をアライメントに用いることで性能の向上が確認できた。

研究成果の概要（英文）：The porous of this research is to improve word alignment accuracy in parallel corpus. In this research, not only word information, but also part-of-speech information and sentence structure are used. Semi-supervised approach is used for training, since it is difficult to additional information to all of sentence in corpus. For Japanese, English, and Chinese parallel corpus, semi-supervised aliment method using POS tag, and meaning tag for proper noun is conducted, and its effectiveness is confirmed. Next, sentence structure information is used for alignment, and its effectiveness is also confirmed.

交付決定額

（金額単位：円）

	直接経費	間接経費	合　計
2008 年度	900,000	270,000	1,170,000
2009 年度	1,000,000	300,000	1,300,000
2010 年度	1,100,000	330,000	1,430,000
年度			
年度			
総　計	3,000,000	900,000	3,900,000

研究分野：総合領域
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、Web を通じてテキストによる情報が爆発的に増大しているにも関わらず、その多くは各国語で書かれたものであることが言語的な障壁となっており、情報の透過的なアクセスのために計算機による機械翻訳の重要性はきわめて高まっている。

大量の言語コーパスを基にした機械翻訳法として現在、統計的学習による統計的機械翻訳が主流となってきている。これは、異言語(例えば、日本語と英語) 間で翻訳関係にある文の大量の対を学習データとして与え、統計的学習により文を異言語に翻訳する翻訳器を自動的に獲得するものであり、その学習のベースになるのが単語アライメントである。

単語アライメントは対訳文間で、どの単語がどの単語に対応しているかを与えるもので、対訳コーパスから機械的に計算される。これは、対訳辞書を自動的に獲得していることに対応しているため、このアライメントの精度は機械翻訳の性能に非常に大きな影響を与える。従って、機械翻訳の性能向上のためには、単語アライメント精度の向上が不可欠な物になる。

２．研究の目的

統計的機械翻訳システムにおいて、その学習モデルの中核となる異言語文間の単語対応付け(アライメント) をとる方法としては、IBM モデル 1～モデル 5 が標準的に用いられている。ただし、対応づけられているのは文のレベルだけであり、文に含まれる単語が異言語間でどう対応しているかの情報は存在しないため、この IBM モデルは実際には次の EM アルゴリズムからなっている。

E ステップ: 異言語の二つの文の間に、隠れた単語間対応(アライメント) を確率的に推定する。
M ステップ: 推定されたアライメントを基に、単語および句の翻訳モデルを学習する。
新しい翻訳モデルを用いて E ステップに戻る。以上を繰り返す。

ここで、単語間対応は日本語の「が」にあたる英語の単語など、 空文字列に対応する場合も含む。これからわかるように、この EM アルゴリズムにおいて、E ステップの単語間対応の学習がアルゴリズムの中核をなしており、統計的機械翻訳はすべてこの対応を基礎としていることがわかる。
しかしながら、このアライメント E ステップは、上の EM アルゴリズムを繰り返しても、現在 60%~70%程度の精度しか得られていない。これは、文中の語自体についての情報がないため、言語的にありえない場合を含め、文中の語が異言語のすべての語と対応する可能性を持つからである。例えば、名詞「首相」が動詞“visited”に対応することは本来ありえないが、これは適切な形態素タグを用いることで回避可能なものと考えられる。
そこで本研究では、生文だけを利用した教師なし学習ではなく、形態素、辞書情報など付与可能な言語情報をすべて統合した、半教師あり学習によるアライメントの学習法を開発し、統計的機械翻訳の精度を根本的な部分から大きく向上させることを目的とする。
使用可能な言語情報には形態素解析や辞書からの情報のほか、構文解析、固有名認識(NER) とその単語翻訳などがあり、これらは近年、多くの言語についてツールにより自動的に付与可能であるが、現在の統計的機械翻訳では使われないままになっている。この他、少量ではあるが人手による直接の単語アライメント正解データも提供されているため、これらを全て訓練データとして使用することで、より正確な単語アライメント推定が可能になると考えられる。従来のアライメント手法では、日本語の単語が英語のすべての単語と対応する可能性を持っているが、言語情報を導入することにより、候補が大きく制限され、より高精度な対応付けが可能になってくると考えられる。

３．研究の方法

最近になっていくつかのアライメント手法が提案されているが、これまでの EM アルゴリズムの中で、アライメントの n-ベスト解を出力して対数線形モデルにより最適なものを選択するもので、直接 EM アルゴリズムの拡張にはなっていない。また、CRF を利用して言語情報を統合するは、訓練データとしてすべてタグが付いている教師ありデータを前提としている。しかしながら、詳細な教師ありデータは一般に少量しか得られないため、タグのない大量の教師なしデータと統合して扱える半教師あり学習が不可欠である。われわれの研究課題は、タグありデータとタグなしデータを統合して、より高精度な単語アライメントを学習する EM アルゴリズムを新しく開発することにある。
これまで、翻訳モデルのトレーニングに標準的に使われているツールキットである GIZA++はタグを一切用いない教師なし学習を基にしているが、我々は可能な言語的情報を全て取り込んだ学習法を提案し、GIZA++を超える学習ソフトウェアの開発を目指している。その際、どのような言語情報が有用であり、どのような言語情報の付与がこれから望まれるかについての知見を蓄積したいと考えている。

４．研究成果

（1）タグ付きコーパス作成のガイドラインの作成とその評価
本研究の目的である半教師あり学習アルゴリズムでは、教師信号であるタグ付きのアライメントコーパスが必要となる。そこで、まずこのコーパス作成のためのガイドラインの作成と、その妥当性の評価を行った。本ガイドラインではアライメントの種類を強対応、弱対応、擬似対応の３種類にわけ、それぞれに対して基準をもうけている。書き言葉コーパス（LDC）と話し言葉コーパス（BTEC）の２種類の中英対訳コーパスに対してアライメントを行い、その評価を行った。

その結果、良好なアライメント精度を得ることができ、正当性を確認できた。

（2）教師あり学習による単語アライメント
半教師あり学習の前段階として、中英対訳コーパスに対して教師ありの条件での単語アライメントを試みた。この時、用いた教師信号は、（１）のガイドラインに従って付与された単語アライメント情報の他、中英の単語間の共起率、中英対訳辞書、文中の相対位置の違い、品詞タグ等である。学習方法としてはコンディショナルランダムフィールド（CRF）を用いている。この教師あり学習によって得られたモデルを用いて単語アライメントを行った結果、従来法である GIZA++ よりも７％低いアライメント誤り率を得ることができ、付加情報の有用性が確認できた。

（3）単語アライメントに対する構文木の利用
対訳関係にある文対の片側の言語文に対してその構文木が与えられた場合、もう片側の言語の単語順序に対して、制約がかかる。たとえば片側の言語文が単語 A、B という列を含み、かつ A と B が部分木をなす場合、もう片側の言語における A と B の対訳語も連続して文中に現れる。この性質を利用して一部の単語アライメント誤りを避けることができる。この手法により翻訳の性能が BLEU 値で１～２％向上し、構文木情報の単語アライメントへの有効性が確認できた。

（４）単言語タグの種類の拡張
２０年度は英語および中国語のパラレルコーパスに対し各単語のタグ情報とアライメント情報を付加した。２１年度は当初２０年度の単語のタグ情報をより詳細なものに発展させる予定であったが、多言語化を先に行うこととした。その理由はタグ情報とアライメント情報は言語に強く依存するため英中２言語では、研究成果が言語対に依存するものかどうかが判断できないためである。そこで、今年度は英中に加え、日本語を加え３各国のタグ、およびアライメント情報付きのパラレルコーパスを整備することとした。

（５）教師あり学習を半教師あり学習に拡張
上記の３各国パラレルコーパスおよび、タグ、アライメント情報なしのパラレルコーパスの両方を用い、半教師あり学習による単語アライメントを試みた。成果としては、日英中いずれの組み合わせにおいてもアライメント精度の向上が確認でき、半教師あり学習の有効性が確認できた。

（６）言語間タグを用いた固有名詞の翻訳
固有名詞等のアライメントに対し、言語ごとに意味情報等を用いてクラス化を行い、クラス情報を用いたアライメントを行った。この結果、固有名詞等の翻訳性能が向上し、有効性が確認できた。

（７）単語アライメントに対する構文木の利用
２０年度は構文構造をアライメントの際の制約として用いたが、アライメントとして許されるかどうかの２値情報としてしか利用していなかった。今年度はこの制約を確率付きのものに拡張することにより、アライメント精度を向上させることができた。

（８）構文木の利用の拡張
２０年度は構文構造をアライメントの際の制約として用い、アライメントとして許されるかどうかの２値情報として利用した。２１年度はこの制約を前後のタグに依存した確率付きのものに拡張することにより、アライメント精度を向上させることができた。本年度は、さらに日本語特有の係り受けにおける制約である Head-Final 制約を英日翻訳に導入することにより、翻訳元言語である英語の単語の位置が日本語では入れ替わるかどうかの制約に用いた。日本語では係り受け関係における係り元と係り先の位置関係において、必ず係り先の位置が後ろになるという制約がある。一方英語ではこのような制約はない。従って、英語において係り先の位置が後ろであれば、日本語と同じ語順になり、前であれば語順が逆転すると言う制約を導入することができる。

（９）単語アライメント評価コーパスの作成
単語のアライメントは内用語間では対応が明確であることが多いが、機能語間では必ずしも明快とは言えない場合が多い。従って、何らかの基準で機能語間のアライメントを定義しなければ、アライメント精度の評価がうまくできない。このため、本年度はアライメントや統計翻訳で最も広く用いられている評価セットである MT08 の英中評価コーパスに対し、アライメントを行い、標準となるアライメント精度評価コーパスとして用いることができるようにした。これにより、他組織における研究結果と公平な評価を行うことができるようになり、今後の研究の大きな助けになると考えられる。

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　３件）

①HASHIMOTO Kei, YAMAMOTO Hirofumi,

OKUMA Hideo, SUMITA Eiichiro, TOKUDAKeiichi, “A Reordering Model Using a Source-Side Parse-Tree for Statistical Machine Translation,” IEICE transactions on information and systems 92(12), 査読あり,2386-2393, 2009-12-01

②YAMAMOTO Hirofumi, OKUMA Hideo, SUMITA Eiichiro, “Imposing Constraints from the Source Tree on ITG Constraints for SMT,” EICE transactions on information and systems 92(9), 査読あり,1762-1770, 2009-09-01

③Chooi-Ling Goh, Eiichiro Sumita, “A Feature-rich Supervised Word Alignment Model for Phrase-based Statistical Machine Translation,” International Journal of Asian Language Processing, 査読あり,Vol. 19, No.3, pp. 109～125, 2009

〔学会発表〕（計　5 件）

① ゴー・チュイリン, 隅田英一郎, “Supervised Word Alignment for Phrase-based Statistical Machine Translation,” 言語処理学会第 15 回年次大会論文集, 鳥取大学,pp.873-876, 2009.

②Hongmei Zhao, Qun Liu, Ruiqiang Zhang, Yajuan Lv, 隅田英一郎, ゴー・チュイリン, "Guidelines for Chinese-English Word Alignment," CWMT'2008 論文集, 北京（中国）,pp.153-163, 2008.

③山本博史,大熊英男, 隅田英一郎, “Imposing Constraints from the Source Tree on ITG Constraints for SMT,” ACL-08: HLT Second Workshop on Syntax and Structure in Statistical Translation (SSST-2),オハイオ（米国）2008

④吉崎大輔,山本博史,大熊英男,匂坂芳典,” 統計的機械翻訳における未登録語のグループ化による翻訳,” 言語処理学会 第16回年次大会, 東京大学,２０１０年３月１０日

⑤西村拓哉, 山本博史,大熊英男,村上仁一,” 英日SMTへのHead-Final制約の導入,” 言語処理学会第17回年次大会,豊橋技術科学大学 ,2011年3月8日

６．研究組織
(1)研究代表者
山本　博史　（YAMAMOTO HIROFUMI）
近畿大学・理工学部・教授
研究者番号：00395013

(2)研究分担者
（　　）

研究者番号：

(3)連携研究者
隅田　英一郎　（SUMITA EIICHIRO）
情報通信研究機構
研究者番号：90395020

安田　圭志　（YASUDA KEIJI）
情報通信研究機構
研究者番号：50395018

ゴー　チュイリン（GHOOI-LING GOH）
情報通信研究機構
研究者番号：90531616