科学研究費助成事業　　研究成果報告書

平成　２６　年　　５　月　　９　日現在

機関番号：３２６８９
研究種目：基盤研究(C)
研究期間：2011～2013
課題番号：２３５００１８７

研究課題名（和文）統計・用例機械翻訳のためのアラインメント向上と多言語文法パターン公開

研究課題名（英文）Improvement of alignment for statistical and example-based machine translation and release of multilingual syntactic patterns

研究代表者
ＬＥＰＡＧＥ　ＹＶＥＳ（LEPAGE, Yves）

早稲田大学・理工学術院・教授

研究者番号：７０５７３６０８

交付決定額（研究期間全体）：（直接経費）　　3,900,000円、（間接経費）　　1,170,000円

研究成果の概要（和文）：従来機械翻訳システムの翻訳知識は翻訳テーブルにある。翻訳テーブルとは、二カ国語辞書に似たようなものであり、通常の辞書より長い見出しを持ち、その見出しの確率等を表す数値をも含めるものである。翻訳テーブルは文部分的アライナーというツールにより自動的に生成される。本研究では先行研究で提案した文部分的アライナー手法の向上ができた。以前より長い見出しを出力し、特定の場合では最高技術水準の翻訳品質を得られることを示した。また、出力された翻訳テーブルを一部公開した：Europarlバージョン３の１１カ国語の共通部分の全言語対を様々な実験設定で得られた翻訳テーブルである。

研究成果の概要（英文）：Current machine translation systems rely on translation tables to translate. Translation tables look like usual dictionaries, but contain longer entries, with numerical values to assess the reliability of the entries. Translation tables are extracted automatically from translated texts by tools called subsentential aligners. In earlier research, we proposed a new subsentential aligner. It is simpler and faster than state-of-the-art tools. But it has lower quality scores in some translation tasks because its entries are not long enough. The goal of the project was to improve the tool so as to achieve as good results as the state-of-the-art tools. This has been achieved in various cases. Many of the translation tables output during the project have been made freely available to the community through a web site (all language pairs between the 11 languages on the common part of the Europarl corpus version 3).

研究分野：総合領域

科研費の分科・細目：情報学・知能情報額

キーワード：自然言語処理　機械翻訳　翻訳テーブル　アラインメント

１．研究開始当初の背景
現在統計機械翻訳システムの翻訳知識は主に翻訳テーブルにある。翻訳テーブルを文部分的アライナーにより生成される。先行研究成果として最高技術水準より簡単で速い文部分的アライナー手法（Anymalign）を開発した。しかし、学習データが２０万文以上の場合、統計翻訳最高技術水準より、翻訳品質が低く、翻訳テーブルの見出しは短いと分析された。

開発した文部分的アライナー手法はサンプリングに基づく手法であるため、anytimeの特徴を持つ。即ち、まず短い時間で近似な解を出力し、その後、近似を最小化の計算によって、徐々に解を改善する手法である。使用者は計算の課程を割り込んだ時間の長さによってより正しい解を得ることができる。翻訳テーブルの生成の場合では、見出しの数と属性の近似は時間にそって向上することである。

２．研究の目的
本研究では開発した文部分的アライナーの精度を最高技術の統計機械翻訳の水準まで向上させる目的であった。また、開発した文部分的アライナーの特徴を利用し、新しい翻訳テーブルの生成手法と使用の検討をする目的であった。具体的に次の検討をする予定だった。

(1) 特定のテストデータ（翻訳したい文に対して）**特定翻訳テーブルの生成**の可能性の検討。

(2) 開発した文部分的アライナーの使用し、**より長い見出しの生成**の可能性の検討。

(3) 開発した文部分的アライナーの使用し、**統計機械翻訳の階層モデル**のための翻訳テーブルの生成の可能性の検討。

(4) 開発した文部分的アライナーの使用で生成された翻訳テーブルを本研究室で開発されている**用例翻訳システム**での使用の検討。

３．研究の方法
(1) 全ての実験を比較できるようにするため、全ての実験で Europarl バージョン３の１１カ国語の共通部分を使用した。それぞれの言語で同じ意味を持っているデータを使用したため、行なった実験結果の差異は言語対の特徴を反映すると解釈できる。共通部分を抽出するため、英語に基づき、それぞれ残りの１０カ国語の対訳関係を持つ文を選択した。１０カ国語全てで存在する文の共通部分はおよそ３８万文に及び、学習データとして 347,614 文、チュニングデータは 500 文、テストデータは 38,123 文とした。

(2) 基本比較のため、ベースライン実験を行なった。そのため、EuroMatrix と呼ばれた Edinburgh 大学で以前に発表された実験を再現した。最高技術水準の統計翻訳ツールを使用し､全ての 110 言語対での翻訳実験を行い、その翻訳品質を従来翻訳品実尺度 BLEU で測定した。

(3) 本研究では、開発した文部分的アライナーを修正したり、前処理のプログラムを使用したり、アラインメント結果を組み合わせたりした。

①　修正としては新しいオプションを二つ追加した：１）特定アラインメント数により停止２）特定した長さ以下の単語列のアラインメント。
②　前処理として同じ長さを持つ単語列のグループの自動生成。開発した文部分的アライナーを利用し、用例翻訳品質の向上を目指し、「可切性」という分かち書きの手法と最良対訳に基づいた二カ国語同時構文解析の手法を導入し、実装した。
③　アラインメントの組み合わせとしは以前に開発した文部分的アライナーの機能であった。

また、一つの翻訳テーブルの内容の解析と二つの翻訳テーブルの比較のために複数の解析ツールを開発した。

(4) 実験をより簡単に行なうため、本研究の予算で本研究室のネットワークを再建した。実験用コンピュータを設定し、扱い安さを向上した。また、データと成果の公開のためサーバを購入した。

４．研究成果
(1) **特定翻訳テーブルの生成。**翻訳したいデータに対して特定翻訳テーブルを生成するため、開発した文部分的アライナーを変更した。翻訳したいデータで出現する単語の出現頻度によって学習データのサンプリング重みを計算し、その重みに基づいてサンプリングを行ないながら翻訳テーブルを生成した。変更なしと変更されたバージョンの比較の結果、同じ翻訳品質を得るため、変更されたバージョンが２倍ほど速い。また、翻訳テーブルのサイズは平均的に８０％減らすことができた。差異のない翻訳品質は翻訳テーブルの属性で説明できる。同じ翻訳テーブル見出しの属性の差異を計算し、平均と標準偏差を比較した結果、語彙重みはほとんど同じ、翻訳確率は５％だけ違う。

(2) **より長い見出しの生成**の研究では、翻訳テーブルを一度に生成するのではなく、先行研究で使われた手法を適用し、同じ長さでの単語列の翻訳テーブルを個別生成する手法を提案した。開発した文部分的アライナーは単単語をアラインメントタスクで最高技術であると以前に証明されたため、同じ長さ

を持つ単語列を単単語として文部分的アライナーに扱わせることで、より効率のよい翻訳テーブルの生成が考えられる。個別に生成された翻訳テーブルを組み合わせることによって全体の翻訳テーブルを合成できる。以上の手法を使用した結果、二つのやり方で最高技術の翻訳品質まで向上することができた。以上の手法は文部分的アライナーに新しいオプションとして追加された（-i オプション、 http://anymalign.limsi.fr/のページを参照）.

① 最高技術水準の統計翻訳ツールの翻訳テーブルと上記で提案した手法で出力された翻訳テーブルを組み合わせることによって、最高技術水準以上の翻訳品質が得られた。さらに、プリューニングを導入することによって、より良い結果が得られた。

② それぞれの長さの翻訳テーブルは anytime の特徴を利用して特定時間で出力する。ここで、三つの時間の分布関数の検討をした。一つ目は全ての長さに対して同じ時間を与える関数、二つ目は源言語と目的言語での長さの差異による正規分布、三つ目はその長さの差異と単語列の長さによる二変量正規分布。以上の三つの手法では、異なる単語列の長さのアラインメントの可能性もあり、より長い見出しの出力もできた。その結果、Europarl 翻訳タスクで翻訳品質が向上した。また、出力された翻訳テーブルを一部公開した。

③ 最良結果としては、一番簡単（ポルトガル語・スペイン語）、一番普通（英語・フランス語）、一番困難（フィンランド語・英語）なEuroparlの言語対の翻訳で、Anymalignで得られた単語列のアラインメントと最良対訳に基づいた二カ国語同時構文解析の組み合わせで最高技術標準より良い翻訳品質を得ることができた（EAMT2102国際会議に発表済）。翻訳テーブルの解析では、少なくとも二つの利点が考察できる。一つは、開発した文部分的アライナーAnymalignだけの使用と比べると、上記で提案した手法では、翻訳テーブルは平均約３倍大きくなった。もう一つは、最高技術標準の翻訳テーブルと比べると、半分の大きさである。Anymalignだけの使用では、翻訳テーブルの見出しの長さは最高技術標準のものより短い。それに対して、上記で提案した手法では、ほぼの同じ長さになった。従って、二カ国語同時構文解析によって、より長い翻訳テーブル見出しの出力ができると示した。扱い易さを目指し、実装を繰り返した結果、二カ国語同時構文解析の手法の加速化ができた。プログラム実行の解析に基づき、基本演算量を減少し、２年前のプログラムに比べて平均約５０倍速くなった。マルチプルロセッシングの導入によってさらなる加速化もできた。その加速化によってこの手法は最高技術水準のツールと比べて時間的に競争力があるといえる。

(3) **統計機械翻訳階層モデル**では、翻訳見出しがルールのようなものになり、変数を含めるものである。最高技術水準ツール Moses のルールでは変数の数が二つ以下である。開発した文部分的アライナーの基本機能として不連続な見出しの出力が可能である。ルールに変換するために特別なツールを設計し、実装した。 源言語と目的語に現れる変数の対応関係を計算は学習データに基づいて行なわなければならないため、計算量の困難があった。解析のため、パターンの変数の平均と標準偏差を測定するツールを開発し、サンプリングで計算量の減少ができた。サンプリングしても完全な計算と比べて、ルールのエラーは３％だけがあり、翻訳品質に影響はないと示した。

(4) **用例翻訳システム**の研究では、先行の研究で提案された単言語の分かち書きと、上記で提案した二カ国語同時構文解析の検討をした。本研究室で開発中の用例翻訳エンジンに対して、統計翻訳のための翻訳テーブルの見出しは短かすぎるため、翻訳は不可能であることを予備実験で証明した。概して翻訳テーブル見出しは文の長さまでのものが必要となる。先行研究で提案された手法は文中の最弱点を位置づけることによって、分かち書きを行なう「加切性」という手法を二つの設定で検討した。一つは独立に源言語と目的言語の文の分かち書きを行なう手法、もう一つは、源言語の可切性で得られた分かち書きの結果に基づいて、目的言語の分かち書きを行なうことである。さらに上記で提案した二カ国語同時構文解析も検討した。 三つ目の手法の方を期待していたのにたいして、二つ目の手法の方が良い結果が得られたが、最高技術標準の翻訳品質には及ばない。

(5) **研究成果の普及**：主な実験設定で得られた翻訳テーブルとその翻訳品質結果をウェブサイトで公開した。110 言語対 x 15 実験設定 x 500Mb（平均）のデータの量である。

lang	da	de	el	en	es	fi	fr	it	nl	pt	sv
da		16.13	15.76	25.44	19.98	10.27	20.15	17.21	18.8	18.58	26.01
de	17.88		15.34	20.03	18.48	8.44	18.38	15.93	20.18	17.05	15.07
el	16.66	14.2		25.21	25.18	8.83	24.26	21.87	16.87	23.58	16.01
en	21.9	15.52	20.65		25.84	10.1	22.7	22.06	20.06	23.45	25.84
es	16.95	13.86	20.62	26.38		8.64	31.28	27.87	17.42	31.93	16.82
fi	14.61	11.2	12.53	17.24	14.84		12.73	13.56	12.77	13.75	13.46
fr	16.79	14.02	19.09	24.82	27.39	7.94		27.04	17.23	27.83	15.86
it	16.18	13.73	19.46	24.49	30.43	8.5	30.62		16.73	28.48	15.38
nl	17.58	17.42	14.6	22.07	18.74	7.97	19.01	16.66		17.1	15.55
pt	16.51	13.8	20.33	24.94	32.84	8.68	31.78	27.87	16.87		15.76
sv	27.23	15.45	16.99	26.68	20.52	9.97	20.84	17.28	18.02	18.79	

【公開された110言語対の実験結果の例】特定スコアをクリックすると翻訳テーブルをダウンロードできる。

本研究が始まる前の開発した分部分的アライナーだけの使用に比べると、本研究で得られた結果はそれより品質の高い結果を得た。複数の場合では、最高技術水準の翻訳品質に近づき、特定場合でも、いくつかはより良い結果を得ることができた。

(6) **国際的承認として、**統計翻訳の分野で先頭を歩いているhttp://www.statmt.orgのサイトで開発した文部分的アライナーの参照が数箇所あり、特にEAMT2012国際会議に採択された論文に参照が見られる（http://www.statmt.org/survey/Topic/WordAlignment）。

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１件）

① J. Luo and Y. Lepage. An investigation of the sampling-based alignment method and its contributions. International Journal of Artificial Intelligence & Applications (IJAIA), 4(4):9-19, July 2013. http://www.airccse.org/journal/ijaia/papers/4413ijaia02.pdf

〔学会発表〕（計１２件）

① J. Luo, A. Lardilleux, and Y. Lepage. Exploring N-grams distribution for sampling-based alignment. In Z. Vetulani, editor, Proceedings of the 5th Language & Technology Conference (LTC'11), pages 289-293, Poznan, November 2011. Fundacja uniwersytetu im. Adama Mickiewicza. http://hal.archives-ouvertes.fr/docs/00/65/08/43/PDF/ltc2011_luo.pdf

② J. Luo, A. Lardilleux, and Y. Lepage. Improving sampling-based alignment by investigating the distribution of n-grams in phrase translation tables. In Proceedings of the 25th Pacific Asia Conference on Language Information and Computing (PACLIC 25), pages 150-159, Singapore, December 2011. http://dspace.wul.waseda.ac.jp/dspace/bitstream/2065/34451/1/150.pdf

③ A. Lardilleux, F. Yvon, and Y. Lepage. Hierarchical sub-sentential alignment with Anymalign. In Federico, editor, Proceedings of the 16th annual conference of the European Association for Machine Translation (EAMT 2012), pages 279-286, Trento, May 2012. Fondazione Bruno Kessler. http://aclweb.org/anthology//E/E14/E14-4.pdf

④ J. Lee and Y. Lepage. Fast production of ad hoc translation tables using the sampling-based method. 言語処理学会第18回年次大会発表論文集 809-812 (2012年3月) http://www.anlp.jp/proceedings/annual_meeting/2012/pdf_dir/P2-20.pdf

⑤ J. Luo, J. Sun, and Y. Lepage. Producing translation tables by separate N-grams subtables. In言語処理学会第18回年次大会発表論文集 797-800 (2012 年 3 月) http://www.anlp.jp/proceedings/annual_meeting/2012/pdf_dir/P2-17.pdf

⑥ J. Luo and Y. Lepage. A comparison of association and estimation approaches to alignment in word-to-word translation. In Proceedings of the tenth international Symposium on Natural Language Processing (SNLP 2013), pages 181-186, Phuket, Thailand, October 2013. Phuket. http://saki.siit.tu.ac.th/snlp2013/uploads_final/62__5260c3fb255ca59586ed311216472f96/snlp.pdf

⑦ J. Luo, A. Max, and Y. Lepage. Using the productivity of language is rewarding for small data: Populating SMT phrase table by analogy. In Z. Vetulani, editor, Proceedings of the 6th Language & Technology Conference (LTC'13), pages 147-151, Poznan, December 2013. Fundacja uniwersytetu im. Adama Mickiewicza.

⑧ T. Kimura, Y. Nishikawa, J. Matsuoka, and Y. Lepage. Generation of translation tables adequate for example-based machine translation by analogy. In Proceedings of the 2014 International Conference on Artificial Intelligence and Software Engineering (AISE2014), page: 200-203, Phuket, Thailand, January 2014. DESTech Publications.

⑨ T. Liu and Y. Lepage. Hierarchical statistical machine translation using sampling based alignment. 情報処理学会研究報告. 自然言語処理研究会報告 2014-NL-215(1), 1-5, 2014. https://ipsj.ixsq.nii.ac.jp/ej/?a

ction=pages_view_main&active_action=repository_view_main_item_detail&item_id=98182&item_no=1&page_id=13&block_id=8

⑩ S. Zhang, J. Luo, and Y. Lepage. Improving N-gram distribution for sampling-based alignment by extraction of longer N-grams. 情報処理学会研究報告. 自然言語処理研究会報告 2014-NL-215(3), 1-5, 2014. http://ci.nii.ac.jp/naid/110009659642

⑪ 西川裕介 木村竜矢 松岡仁 ルパージュ・イヴ 単言語または二言語の分割性による類推翻訳の検討 言語処理学会第 20 回年次大会発表論文集 598-601 (2014 年 3 月) http://kmcs.nii.ac.jp/nlp_annual/nlp2014-P6-06/ye=2014&ys=2014&tf=2&o=140

⑫ J. Luo and Y. Lepage. Production of phrase tables in 11 European languages using an improved sub-sentential aligner. In Proceedings of the 9th edition of the Language Resources and Evaluation Conference (LREC 2014), (to appear), Reykjavik, Iceland, May 2014. ELRA.

〔図書〕(計１件)

① J. Luo, A. Lardilleux, and Y. Lepage. Improving the distribution of N-grams in phrase tables obtained by the sampling-based method. Lecture Notes in Artificial Intelligence, 2013, 12 pages (to appear).

〔その他〕
ホームページ等
http://133.9.48.109/index.php/kakenhi

６．研究組織
(1)研究代表者
ルパージュ・イヴ (LEPAGE, Yves)
早稲田大学・大学院情報生産システム研究科・教授
研究者番号：70573608