F-06

文クラスタを利用したドメイン依存統計翻訳

Domain Dependent Statistical Machine Translation Using Sentence Clustering

塩田 好†　　山本 博史‡
Yoshimi Shiota　　Hirofumi Yamamoto

1. 序論

近年、機械翻訳の手法として主流となりつつある統計翻訳は、翻訳元言語と翻訳先言語の対訳文を大量に集めた対訳コーパスから、自動的にシステムを構築するものであり、専門家の知識や対訳辞書を必要としないという利点がある。統計翻訳は文字通り統計に基づく手法であるため、統計量をとる対象である対訳コーパスの量が多いほうが望ましい。また、対訳コーパスから得られた統計量に基づいて翻訳を行うため、翻訳対象の文の統計的性格が対訳コーパスと同じである事が前提となる。

大量の対訳コーパスを取得するためには、どうしても特定の分野(ドメイン)によらず収集する必要がある。従って、特定のドメインの文を翻訳する場合には、翻訳対象の統計的性格と対訳コーパスの統計的性格が同じであるという前提が崩れる可能性がある。

そこで本研究では、大量の対訳コーパスを予め統計的性格が似たような文が集まるようクラスタリングし、そのクラスタをドメインとして見なし、翻訳対象の文と統計的性格が最も近いドメインを用いて翻訳を行うことで、翻訳対象の文と対訳コーパスの間の統計的性格の差を小さくする。そして、翻訳精度が最適となる分割パターンを模索する。

2. 統計翻訳

統計翻訳は、翻訳文fに対して、最も確率が高くなる翻訳文eを見つける問題であり、次式で表すことができる。

$$\arg_e max\ P(e|f) = \arg_e max\ P(f|e)P(e)$$

このうち、右辺の$P(f|e)$が翻訳モデル、$P(e)$が言語モデルと呼ばれるものであり、eに対しなるべく高い確率を与えるものが望ましく、これはすなわち、翻訳モデル、言語モデルのエントロピーをなるべく少なくすれば良い事を意味している。

3. 提案手法

本手法は文クラスタリングとクラスタベース翻訳の２つの手順で行われる。翻訳手順の概要を図 1 に示す。

文クラスタリングの手順で、学習用対訳コーパスをクラスタリングし対訳サブコーパスを作成し、そこから得た各クラスタがドメインとなる。さらにドメインごとにモデルが作成される。クラスタベース翻訳では、入力文のドメインを推定し、最適なモデルを用いて翻訳する。

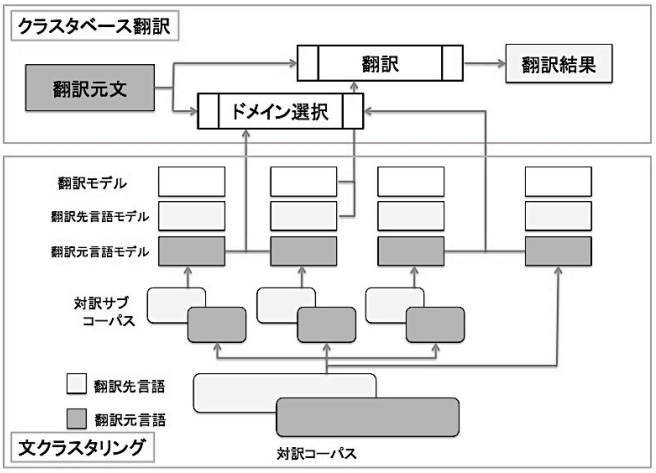

図 1.翻訳手順の概要

ドメインを動的に変化させることを、確率変数 C を用いて以下の式で表すことができる。

$$\arg_{e,C} max\ P(e,C|f) = \arg_{e,C} max\ P(C|f)P(e|f,C)$$

右辺の$P(e|f,C)$がドメイン依存翻訳モデル、$P(C|f)$がドメイン依存言語モデルとなり、fに対しなるべく高い確率を示すドメインを使用する。以下に処理の詳細を示す。

3.1 文クラスタリング

クラスタリングは対訳文を統計的性格の似た者同士に分ける事が目的である。対訳コーパスをエントロピーに基づいてクラスタリングし、そこから得られたクラスタをサブコーパス、つまりドメインと見なす。文クラスタリングの手順は以下の通りである。下記では分割するクラスタ数は N 個とする。

1. 全ての対訳文をランダムに N クラスタに分配する。
2. あるクラスタの 1 文を選択し、他のクラスタに移動したと仮定し、エントロピーを計算する。
3. 2.の処理を全クラスタに対して行い、エントロピーが最も小さくなったクラスタに対訳文を移動させる。
4. 2,3 の手順を全ての文について行う。
5. エントロピーの減少量を測り、それが閾値を下回るまで 2.以降の手順を繰り返す。
6. 全クラスタと対訳文全体で N+ 1 個の翻訳元、翻訳先の言語モデル、翻訳モデルを作成する。

†近畿大学大学院総合理工学研究科，Graduate School of Science and Engineering. Kinki University
‡ 近畿大学理工学部，Department of Science and Engineering. Kinki University

3.2 パラメータチューニング

各ドメインの統計的性格の似たデータを使用してパラメータチューニングを行い、ドメインの統計的性格を生かすように調整する。手順は以下の通りである。

1. 文クラスタリングから得られた N+1 個の翻訳元言語の言語モデルを使用して、デベロップメント文を最も確率の高くなるドメインに配分し、デベロップメント文を N+1 個に分割する。
2. 非分割の対訳文全体には分割する前のデベロップメント文全体を、各ドメインには各ドメインに配分されたデベロップメント文を用いて、パラメータチューニングを行う。

3.3 クラスタベース翻訳

翻訳する文の統計的性格を測り、動的に統計的性格が最も近い翻訳モデル、言語モデルを用いて翻訳する。翻訳手順は以下の通りである。

1. 各クラスタ、および、全ての文を学習データとした N+1 個の翻訳モデル、および翻訳元、翻訳先の言語の言語モデルを学習する。
2. N+1 個の翻訳元言語モデルを用いて、翻訳対象文に対し、エントロピーを求める。
3. 最も近いエントロピーを示したクラスタ(または全文)から作成された翻訳先言語モデル、翻訳モデルを用いて翻訳を行う。
4.

4. 実験と結果

以下に実験条件と結果を示す。

4.1 実験条件

実験に用いた学習用データは英日特許文約 180 万文対で、2 個、4 個、8 個のクラスタに分割し、モデルを生成し、計 15 個のドメインを使用した。ここで特許文を使用した理由として、特許文には専門用語を数多く含み、特定分野の文が多く含まれていると考えられるためである。文クラスタリングにおいて、閾値を定義する代わりに、ループ回数を事前に 20 回として決定し、文クラスタリングを行うこととした。パラメータチューニングに用いるデベロップメント文は学習文に含まれない特許文である 915 文を用いた。

翻訳方向は英語から日本語の英日翻訳であり、翻訳モデルの学習には moses[1]、言語モデルの学習には srilm[2]、パラメータチューニングには moses に含まれる mert-moses-new.pl を用いた。評価には学習文、パラメータチューニング文に含まれない特許文である 1381 文を用い、評価尺度は blue[3]を用いた。

4.2 予備実験

まず、2 分割、4 分割、8 分割のクラスタの有効性の確認を行った。各分割数に対し、評価文がそのクラスタに属すると判断される数、すなわち、そのクラスタの言語モデルが最も低いエントロピーを示す文数の 1 クラスタ当りの平均文数は、非分割が 123 文、2 分割が 42 文、4 分割が 35 文、8 分割が 130 文となった。このことから、2 分割、4 分割のクラスタは有効ではないと判断でき、実験には非分割と 8 分割のクラスタのみを用いることとした。

4.3 実験結果

8 クラスタへの分割の結果、各クラスタに含まれる文数はどのクラスタも概ね等しかった。次に、このクラスタで生成した翻訳元言語の言語モデルを用いてデベロップメント文の分割を行った結果、どのクラスタに対しても約 100 文が割り当てられ、これを用いてクラスタごとにパラメータチューニングを行った。なお、非分割に対しては全てのデベロップメント文を用いてパラメータチューニングを行った。

次に、評価文 1381 文を言語モデルで分割した結果、各ドメインに割り当てられた文数を表 1 に示す。なお、用いた言語モデルは、翻訳元言語、翻訳先言語共に 3-gram である。

この手順に基づいて翻訳を行った結果、blue 値は 33.20 となった。従来のクラスタリングを行わなかった場合の blue 値が 28.71 であったことに対し、提案手法における blue 値の向上は 4.49 と大きな性能向上を示し、有効性が確認できた。

表 1.各ドメインの評価文の割当数

分割数	1	8							
No.	1	1	2	3	4	5	6	7	8
割当数	225	165	220	280	62	56	115	100	158

5. 結論と課題

本研究では文クラスタリングを用いたドメイン依存統計翻訳を行い、有効性が確認できた。また分割数が大きくなるにつれ、統計的性格がドメインごとに細分化され、翻訳精度に効力を与える。

しかし今回の実験では、分割数は 2 分割、4 分割、8 分割の 3 種類と少なく、かつ分割数が大きくなるにつれ、ドメインの選択される回数が多いことから、さらに分割数を増やすことや、統計的性格の近いドメインの結合を行うことによる、性能の変化を調べることが求められる。

参考文献

[1]. Moses, 2014/07/15,http://www.statmt.org/

[2]. Srilm-SRI Speech Technology and Research Laboratory, 2014/07/15, http://www.speech.sri.com

[3]. Papineni Kishore, Salim Roukos, Todd Ward, Wei-Jing Zhu: BLEU: a method for automatic evaluation of machine translation", 40th Annual meeting of the Association for Computational Linguistics pp. 311-318, 2002.

[4]. Franz Josef och, Minimum Error Rate Training in Statistical Machine Translation, Association for Computational Linguistics 2003, pp160-167, 2003