素性空間拡張とコーパス結合を併用した統計翻訳のマルチドメイン適応

今村 賢治　　隅田 英一郎
国立研究開発法人 情報通信研究機構
{kenji.imamura,eiichiro.sumita}@nict.go.jp

1 はじめに

さまざまな種類のテキストや音声認識結果が機械翻訳されるようになってきている．しかし，すべてのドメインのデータにおいて，適切に翻訳できる機械翻訳器の実現は困難であり，翻訳対象ドメインを絞りこむ必要がある．

対象ドメインの翻訳品質を確実に向上させるには，学習データ（対訳文）を大量に収集し，翻訳器を訓練することである．しかし，多数のドメインについて，対訳文を大量に収集することはコスト的に困難であるため，他のドメインの学習データを用いて対象ドメインの翻訳品質を向上させるドメイン適応技術が研究されている．

本稿では，複数のドメインに同時に適応させる，機械翻訳のマルチドメイン適応方式を提案する．本稿の提案方式は，基本的には機械学習の分野で用いられている素性空間拡張法 (Daumé, 2007) を機械翻訳に適用したものであるが，機械翻訳における従来のドメイン適応方法の一つであるコーパス結合方式も併用する．

実験では，対象ドメインの訓練コーパスが小さい場合，対象ドメインのデータのみを用いる単独モデルに比べ，翻訳品質は大幅に高く，訓練コーパスが大きい場合でも，単独モデルの翻訳品質と同等となった．

2 機械翻訳のドメイン適応

コーパス結合　最もシンプルなベースラインとして用いられている方法は，翻訳対象ドメイン（IN ドメイン）と他のドメイン（OUT ドメイン）のデータを結合して学習し，1 つのモデルを構築する方法である（本稿ではコーパス結合方式と呼ぶ）．

一般的な機械学習では，結合されたコーパスで学習したモデルは，IN ドメイン，OUT ドメイン双方の中間的性質を持つため，その精度も IN データのみ，OUT データのみで学習されたモデル（ドメイン依存モデル，または単独モデルと呼ぶ）の中間の精度になることが多い．機械翻訳の場合，コーパスを結合することにより，カバーする語彙が増加するため，未知語が減少し，単独モデルより翻訳品質が向上する場合もある．最終的に翻訳品質が向上するか否かは，パラメータの精度低下と未知語の減少のトレードオフになる．

線形補間，対数線形補間　統計翻訳では，翻訳で使用するサブモデル（フレーズテーブル，言語モデル，並び替えモデルなど）の返す値（素性関数値）を線形または対数線形結合して，翻訳文の尤度を算出する．ドメイン適応では，ドメイン毎に素性関数の重みを切り替えることで，ドメイン依存の翻訳を生成する方法がある (Foster and Kuhn, 2007)．重みの推定には，誤り率最小訓練法 (MERT) などが用いられている．

素性空間拡張法 (Daumé, 2007) は，翻訳に限らず，機械学習全般に使われるドメイン適応方式で，素性関数の重みをドメイン毎に最適化する（3.1 節参照）．Clark et al. (2012) は，これを対数線形補間方式の一種として翻訳に適用し，効果があったと報告している．なお，彼らは単一のモデルを用いており，素性関数の重みだけをドメイン適応させている．

fill-up 法　fill-up 法 (Bisazza et al., 2011) は，素性関数値を適応させる方法の一種である．これはフレーズテーブルから翻訳候補を取得する際，IN ドメインの単独モデルにフレーズが存在する場合はその素性関数値を使用，存在しない場合は OUT ドメインの単独モデルからフレーズを取得し，その値を使用する．素性関数値のみでなく，フレーズ候補も変更するため，未知語は減少する．

提案方式の位置づけ　本稿の提案方式は，単独モデルおよびコーパス結合方式を用いて作成したモデルを，対数線形補間で最適化する方法と位置付けられ，未知語が減少とモデルの重みのドメイン最適化が同時にできることが期待される．

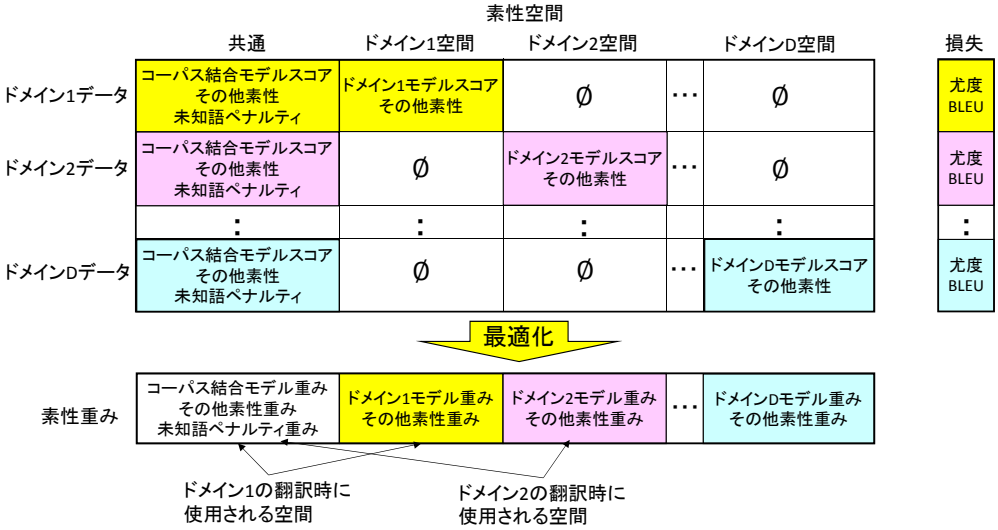

図 1: 素性空間拡張とコーパス結合モデルの併用

3 マルチドメイン適応方式

3.1 素性空間拡張法

素性空間拡張法 (Daumé, 2007) は，機械学習における素性の重み最適化に用いられる方式である．素性空間を共通，OUT ドメイン，IN ドメインに分割し，素性を，それが由来するドメインごとに異なる素性空間に配置し，全体を最適化する．

通常は，OUT ドメインを IN ドメインに最適化するために使用されるが，素性空間拡張法では，OUT ドメインと IN ドメインを同等に扱っており，容易に D ドメインに拡張することができる．その場合，素性空間は共通，ドメイン 1, ... ドメイン D のように，$D+1$ 空間に分割される（図 1）．

統計翻訳の場合，各空間に配置する素性は翻訳モデルスコア，言語モデルスコアなどの素性関数値であるので，これらの値で対数線形モデルを構成し，開発セットを用いて，全体を最適化することになる．

3.2 コーパス結合モデルとドメイン依存モデルの導入

機械翻訳では，モデルの重みだけでなく，モデルそのものが返すスコア（素性関数値）も翻訳品質に影響を与える．この素性関数値に影響する要素としては，関数値そのもの（インスタンス）の精度と，インスタンスのカバレッジが考えられる．

本稿で対象とする句に基づく統計翻訳 (PBSMT) で使われるモデルには，フレーズテーブル，語彙化並び替えモデル，言語モデルなどがあるが，このうちフレーズテーブルは，句の翻訳対を含み，翻訳仮説生成にも使われるため，カバレッジが十分にないと未知語が頻発することになる．

そこで本稿では，インスタンスのカバレッジを確保するためのコーパス結合モデルと，素性関数値の精度を保証するためのドメイン依存モデルの両方を導入する．具体的には，

- フレーズテーブル，語彙化並び替えモデル，言語モデルなどの各モデルについて，ドメイン依存モデルと，コーパス結合モデルを作成しておく．
- 共通空間にはコーパス結合モデルのスコアを配置し，各ドメインの空間にはドメイン依存モデルのスコアを配置し，最適化する（図 1）[1]．
- デコーディングの際は，まず，ドメイン依存モデルとコーパス結合モデルのフレーズテーブルを OR 検索し，翻訳仮説を生成する．探索の際には，共通空間と対象ドメインの空間の素性だけを使って尤度計算をする．

翻訳仮説生成にコーパス結合フレーズテーブルを使用することにより，他のドメインで出現した翻訳対も利用でき，未知語の減少が期待できる．また，インスタンスがドメイン依存モデルに存在している場合，高い精度の素性関数値になることが期待される．

4 節の実験では，Moses ツールキット (Koehn et al., 2007) のデフォルト素性を用いる．これを素性空間拡張すると，共通空間では 15 次元，ドメイン依存空間では各 14 次元となる．

[1] モデルにインスタンスが存在しない場合，今回は素性関数値として-100 を返した．

表 1: コーパスサイズ (文数)

ドメイン	訓練	開発	テスト	英単語数/文
MED	222,777	1,000	1,000	14.0
LIVING	986,938	1,800	1,800	15.5
NTCIR	1,348,871	2,000	2,000	31.7
ASPEC	997,181	1,790	1,784	25.7

3.3 最適化

一般的な機械学習における素性空間拡張法の利点の一つは，素性空間を操作しているだけなので，既存の最適化アルゴリズムが使えるという点である．

本稿では，最適化に k ベストバッチ MIRA（以下 kbmira）(Cherry and Foster, 2012) を用いるが，通常の機械学習における最適化と機械翻訳の最適化の大きな相違点は，多くの機械学習の損失関数が，尤度などデコーダが出力するスコアを使用しているのに対して，機械翻訳は BLEU のような，翻訳文の自動評価値を使用する点である．この自動評価値は，翻訳文と参照訳との比較によって算出され，**コーパス単位に**計算される場合が多い．実際，kbmira は開発コーパスの BLEU スコアを損失関数の一部に使用している．つまり，複数ドメインを同時に最適化する場合は，ドメイン毎に BLEU スコアを算出しないと，最適化結果がドメイン最適にならないことを意味している．

上記問題を解決するため，本稿では kbmira を変更する．文献 (Cherry and Foster, 2012) の Algorithm 1 に対する変更点は，以下のとおりである．

1. 処理済み翻訳文の BLEU 統計量（n-gram 一致数など）を保存する変数 BG を，1 つからドメイン数 D 個に拡張する．
2. 各翻訳文の BLEU スコアは，その翻訳文のドメイン d の BG_d から算出する．
3. 素性重みを更新後，その翻訳文の BLEU 統計量を BG_d に追加する．

この変更によって，各ドメイン空間の素性重みは，そのドメインの開発コーパスに最適化される．

4 実験

4.1 実験設定

ドメイン/コーパス　本稿では，英日／日英翻訳を対象に，以下の 4 つのドメインの同時最適化を行う．各ドメインのコーパスサイズを表 1 に示す．なお，訓練文は 80 単語以下のものだけを使用している．

MED: 病院等における医師（スタッフ）と患者の疑似対話のコーパス．内部開発．

LIVING: 外国人が日本に旅行や在留する際の疑似対話コーパス．内部開発．

NTCIR: 特許コーパス．訓練コーパスと開発コーパスは NTCIR-8，テストコーパスは NTCIR-9 のものを使用 [2]．

ASPEC: 科学技術文献コーパス [3]．ASPEC-JE のうち，対訳信頼度の高い 100 万文を使用．

翻訳システム　各コーパスの対訳文は，内部開発の事前並べ替えを適用したのちに使用した．翻訳システムの訓練のうち，フレーズテーブル，語彙化並び替えモデルの学習には Moses をデフォルト設定で使用した．言語モデルは KenLM を用いて 5 グラムモデルを構築した．最適化は 3.3 節で述べたマルチドメイン kbmira を使用した．デコーディングには，内部開発の Moses のクローンデコーダを使用した．デコーダの設定値は Moses のデフォルト値と同じ `phrase_table_limit=20`，`distortion_limit=6`，ビーム幅 200 とした．

評価指標　評価指標には BLEU を使用し，MultEval ツール [4] で有意差検定を行った．危険率は $p < 0.05$ とした．最適化の揺れを吸収するため，5 回最適化を実施し，その平均値を使用した．

比較方式　以下の方式を比較する．

1. **単独モデル**: 各ドメインコーパスだけでモデルを構築，最適化，テストした場合．これをベースラインとして，他の方式と比較する．
2. **コーパス結合**: コーパス結合モデルを使用し，各ドメインの開発コーパスで最適化，テストした場合．
3. **Fill-up 法**: ドメイン適応方式に fill-up 法 (Bisazza et al., 2011) を用いた場合．
4. **素性空間拡張法 1 (Clark)**: 共通空間，ドメイン空間共に，コーパス結合モデルの素性関数を使った素性空間拡張法．Clark et al. (2012) の設定と同じだが，最適化にはマルチドメイン kbmira を使用した．
5. **素性空間拡張法 2 (提案法)**: 共通空間にはコーパス結合モデルの素性関数を使用し，ドメイン依存空間ではドメイン依存モデルを使用した素性空間拡張法．

[2] http://research.nii.ac.jp/ntcir/index-ja.html
[3] http://lotus.kuee.kyoto-u.ac.jp/ASPEC/
[4] https://github.com/jhclark/multeval．このツールでは bootstrap resampling を 10,000 回繰り返している．

表 2: 方式別の BLEU スコア

方式	英日翻訳 最適化/テストコーパス MED	LIVING	NTCIR	ASPEC	日英翻訳 最適化/テストコーパス MED	LIVING	NTCIR	ASPEC
単独モデル	23.23	24.56	38.62	32.69	17.38	19.71	33.63	21.75
コーパス結合	22.65(-)	22.99(-)	38.09(-)	30.59(-)	17.07	18.80(-)	33.21(-)	20.41(-)
Fill-up 法	22.42(-)	23.38(-)	38.37(-)	31.50(-)	16.56(-)	19.06(-)	33.14(-)	20.98(-)
素性空間拡張法 1 (Clark)	22.49(-)	22.97(-)	38.09(-)	30.65(-)	16.75(-)	18.95(-)	33.24(-)	20.39(-)
素性空間拡張法 2 (提案法)	23.28	24.04(-)	38.67	32.69	17.06(-)	19.91(+)	33.60	21.65

表 3: 訓練コーパスサイズ別翻訳品質 (MED 英日)

訓練サイズ	単独モデル	素性空間拡張法 2
1k	6.42	16.06 (+)
3k	8.99	16.01 (+)
10k	12.54	17.10 (+)
30k	16.49	18.27 (+)
100k	20.63	21.14 (+)
223k (全訓練文)	23.23	23.28

4.2 翻訳品質

各方式について，英日翻訳および日英翻訳における BLEU スコアを表 2 に示す．なお，表中の (+) は，単独モデル方式をベースラインとしたとき，有意に向上したもの，(-) は有意に悪化したものを表す（$p < 0.05$）．

単独モデルと比較した場合，コーパス結合方式は，翻訳品質も単独モデルより低下する傾向が強かった．素性空間拡張法 1(Clark) でも同様で，コーパス結合モデルだけを使う方式は，単独モデルより翻訳品質が下がった．これはコーパス結合方式は各ドメインが平均化されたモデルが作成されるため，素性関数の精度が落ちたためと，単独モデル自体が比較的大きな対訳コーパスから構築されているため，単独でも翻訳品質が確保できたためと考えられる．Fill-up 法は，コーパス結合方式に比べると翻訳品質は向上する場合が多かったが，単独モデルより悪化した．

素性空間拡張法 2（提案法）は，単独モデルとほぼ同等な翻訳品質となった．最適化された素性重みの精度は，複数ドメインの単独モデルと同等にできた．提案方式はコーパス結合モデルを併用しているので，単独モデルに比べて未知語は減少しているのだが，それが翻訳品質の向上にはつながらなかった．

4.3 シングルドメイン適応としての効果

ドメイン適応が必要となる場面は，新たなドメインデータの翻訳を行わなければならないにも関わらず，十分な量の訓練文が集まらない場合である．本節では，MED 英日翻訳に絞って，訓練コーパスのサイズを変えて翻訳品質を測定する．

表 3 は，単独モデルと素性空間拡張法 2（提案法）を比較した結果である．訓練コーパスのサイズが小さい場合には，明らかに提案法の翻訳品質が勝っている．これは，コーパス結合モデルの併用により未知語が減少した効果が直接表れたものである．

5 おわりに

本稿では，素性空間拡張法を機械翻訳に適用し，マルチドメイン適応を行った．共通空間に配置する素性をコーパス結合モデルから，ドメイン依存空間に配置する素性をドメイン依存モデルから獲得し，マルチドメイン kbmira で最適化した．

結果，特に訓練コーパスが小さい場合に，単独モデルと比べて非常に高い翻訳品質が確保でき，訓練コーパスが大きくなっても，単独モデルとほぼ同等の翻訳品質を確保できることを示した．

謝辞

本研究は総務省の情報通信技術の研究開発「グローバルコミュニケーション計画の推進-多言語音声翻訳技術の研究開発及び社会実証- I. 多言語音声翻訳技術の研究開発」の一環として行われました．

参考文献

Arianna Bisazza, Nick Ruiz, and Marcello Federico. 2011. Fill-up versus interpolation methods for phrase-based SMT adaptation. In *Proc. of IWSLT-2011*.

Colin Cherry and George Foster. 2012. Batch tuning strategies for statistical machine translation. In *Proc. of HLT-NAACL 2012*, pages 427–436.

Jonathan H. Clark, Alon Lavie, and Chris Dyer. 2012. One system, many domains: Open-domain statistical machine translation via feature augmentation. In *Proc. of AMTA 2012*.

Hal Daumé, III. 2007. Frustratingly easy domain adaptation. In *Proc. of ACL-2007*, pages 256–263.

George Foster and Roland Kuhn. 2007. Mixture-model adaptation for SMT. In *Proc. of WMT-2007*, pages 128–135.

Philipp Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin, and E. Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proc. of ACL-2007 Demo and Poster Sessions*, pages 177–180.