国際特許分類を用いた特許文機械翻訳に関する検討

安田 圭志†,‡　　隅田 英一郎†,‡

†情報通信研究機構音声言語グループ
‡ATR 音声言語コミュニケーション研究所

〒619-0288「けいはんな学研都市」光台二丁目 2 番地 2

E-mail:　†{keiji.yasuda, eiichiro.sumita}@nict.go.jp

1．はじめに

近年の研究により，コーパスベース機械翻訳[1]の有効性が示されている．コーパスベース方式の中でも，統計翻訳[2]は，対訳コーパスさえあれば，人手を要すること無しに，自動的に翻訳システムを構築することが可能であるため，システム開発に要する時間とコストの少なさにおいて，大きなメリットを持っている．

統計翻訳では，文単位で対応付けられた対訳コーパスが最も重要な構成要素のひとつであり，得られるシステムの性能は，システム自動学習時に用いる対訳コーパスの質と量とに大きく影響される．そのため，統計翻訳システムの研究開発を行う上では，対訳コーパスの利用可能性について考慮する必要がある．

日英の言語対に限定すれば，特許の分野が最も大規模な対訳コーパスが整備されている．また，特許翻訳を自動化するニーズも非常に高いことから，統計翻訳を適用するには非常に適した分野であると言える．

本研究では，統計翻訳をベースにした特許翻訳システムについて扱う．特に，特許記事においては，発明内容の分野により，用いられる単語が大きく異なるという問題がある．このような問題点を解決するため，国際特許分類（IPC）の情報を用い，各 IPC 毎に学習された IPC 依存モデルと，全てのデータを用いて学習された一般モデルとを補間して用いることにより，訳質の向上を図っている．

以下では，2 で IPC について概説し，4 で提案手法について説明する．次に 4 で実験結果について述べ，最後に 5 で論文を結ぶ．

2．国際特許分類

国際特許分類（IPC）は，以下の 8 つのセクションを頂点とした階層構造を持つ分類方式である．

A：生活必需品
B：処理操作；運輸
C：化学；冶金
D：繊維；紙
E：固定構造物
F：機械工学；証明；加熱；武器；爆破
G：物理学
H：電気

セクションの下には，クラス，サブクラス，グループ，サブグループなどの下位階層があるが，本研究では，セクションの情報のみを用いた．

3．提案手法

図 1 に提案手法の処理の流れを示す．提案手法では，まず，全ての対訳コーパスを用いて，原言語側言語モデル，目的言語側言語モデル，翻訳モデルをそれぞれ学習する．以降，これらのモデルを一般モデルと呼ぶ．

次に，IPC に基づきコーパスを分類し，各セクション毎のサブコーパスを作成する．これらのサブコーパスを用いて，セクションごとに，IPC 非依存モデルの場合と同様，

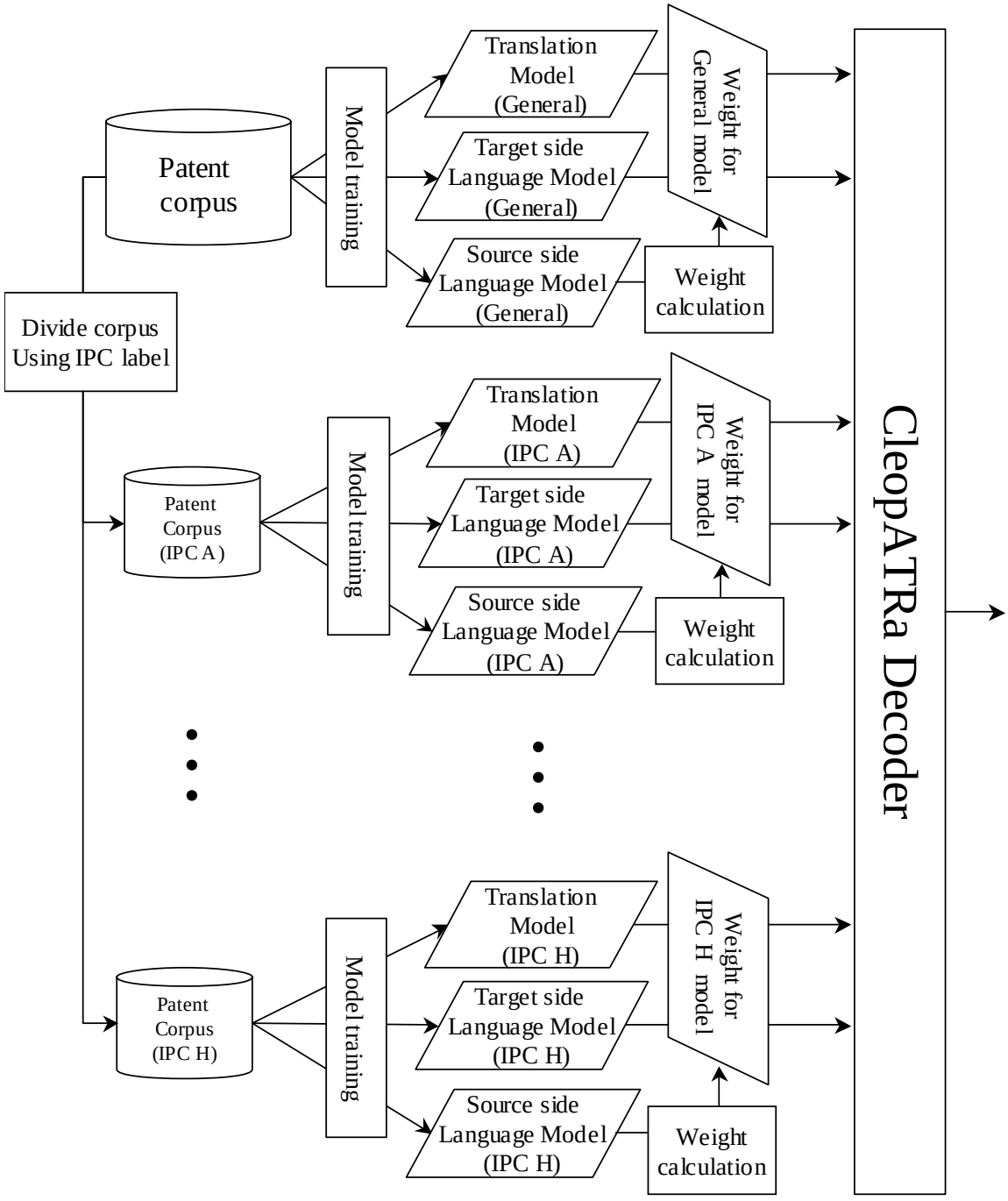

図 1　提案手法の処理の流れ

モデル学習を行う．ここで得られるモデルを IPC 依存モデルと呼ぶ．

通常の統計翻訳では，各種翻訳モデルと，目的言語側言語モデルが必要であり，原言語側言語モデルは必要無い．本研究では，原言語側言語モデルを，一般モデルと複数の IPC 依存モデルとを補間する際の重み計算のために用いる．

ここでまず，従来法の統計翻訳について概説する．従来法の統計翻訳では，次式を用いて，入力文(f)に対する最適な出力文($\hat{e}$)を探索する．

$$\hat{e}(f) = \arg\max_{e} \exp\left\{\sum_{i=1}^{M} \lambda_i \cdot h_i(e, f)\right\}$$

ここで，h(e,f)は言語モデル，翻訳モデルなどの素性関数であり，M は翻訳時に用いる素性関数の数を表す．

提案手法では，入力文毎に，次式を用いて，最適な出力文($\hat{e}$)を探索する．

$$\hat{e}(f) = \arg\max_{e} \exp\left\{\sum_{i=1}^{M} \sum_{j \in IPC} \lambda_{i,j} \cdot \mu_j \cdot h_{i,j}(e, f)\right\}$$

だだし，

$$IPC = \{A, B, C, D, E, F, G, H, General\}$$

である．

また，各セクションの IPC 依存モデルや一般モデルに対する重みである μ_j は，次式により計算する．

$$\mu_j = \frac{P(j \mid S_{input})}{\sum_{k \in IPC} P(k \mid S_{input})} = \frac{P(S_{input} \mid j) \times P(j)}{\sum_{k \in IPC} P(S_{input} \mid k) \times P(k)}$$

$P(S_{input} \mid j)$ ならびに，$P(S_{input} \mid k)$ については，先に述べたように，各入力文の生起確率を IPC 依存の原言語側言語モデルを用いて計算することにより算出する．また，$P(j)$ と $P(k)$ については，各セクションの生起確率であるため，全コーパスの内，今注目しているセクションに属する文の割合を算出することにより計算できる．

4．実験

4.1 実験条件

実験では，NTCIR-7 Patent translation task [3]で提供されたデータを用いている．表 1 にモデル学習に用いた学習セットとパラメータチューニングに用いた開発セットの詳細を示す．

翻訳モデルの学習には MOSES[4]を，言語モデルの学習には SRI language model tool kit[5]を用いた．

評価にはテストセットとして，NTCIR-7 Patent translation task の formal run で使用された 1381 文を用い，リファレンス数 1 の BLEU[6]により自動評価を行っている．

4.2 実験結果

表 2 に英日方向の翻訳実験の実験結果を，表 3 に日英方向の翻訳実験の実験結果をそれぞれ示す．表 2,3 において，3 列目のフィールドは，提案手法を用いたかどうかを表している．このフィールドが No の場合は，一般モデルのみを用いた結果を表している．4 列目のフィールドは，言語モデル学習時に単言語コーパス(表 1 参照)を用いたかどうかを表しており，このフィールドが No の場合は，言語モデル学習時に対訳コーパスの片言語側のデータのみを用いていることを表す．5 列目のフィールドは，目的言語側の言語モデルの n-gram の次数を表している．デコーディング時に必要となるメモリ量の関係で，提案手法を用いる場合は，用いない場合よりも言語モデルの次数を落とし 4 としている．

表 2 を見ると，英日方向の翻訳では，提案手法を用いず単言語コーパスと高次の n-gram を用いた場合 (Run 3) が最も良い結果となっており，次いで提案手法と単言語コーパスの両方を用いた場合(Run 1)となっている．

一方，表 3 を見ると，日英方向の翻訳では，提案手法を用いた場合(Run 1)が最も良い結果となっており，高次の言語モデルや単言語コーパスを用いた場合よりも勝っている．

表 1　学習セットと開発セットの詳細(文数)

	IPC								ALL (General)
	A	B	C	D	E	F	G	H	
Monolingual training set (en)	30.8 M	28.3 M	29.4 M	1.8 M	3.5 M	10.7 M	49.9 M	32.0 M	190.9 M
Monolingual training set (ja)	22.8 M	49.0 M	41.1 M	4.2 M	9.0 M	20.1 M	85.3 M	63.2 M	282.2 M
Training set	58.3 K	271.4 K	41.0 K	10.6 K	6.8 K	161.0 K	1122.1 K	751.9 K	2423.2 K
Dev. set	14	75	16	1	1	54	489	265	915

表 2　英日方向翻訳実験の評価結果

TasK	RUN	IPC-based model	Monolingual corpus	Language model order	BLEU
EJ	1	Yes	Yes	4	29.15
EJ	2	Yes	No	4	29.08
EJ	3	No	Yes	5	**29.22**
EJ	4	No	No	5	29.14

表 3　日英方向翻訳実験の評価結果

TasK	RUN	IPC-based model	Monolingual corpus	Language model order	BLEU
JE	1	Yes	No	4	**24.79**
JE	2	No	Yes	5	23.34
JE	3	No	No	5	24.52

5. まとめと今後の検討課題

IPC を用いた統計翻訳の枠組みを提案し，NTCIR-7 Patent translation task のデータを用いた実験を行った．英日方向では，ベースラインに劣るものの，日英方向では，ベースラインを上回る性能が得られた．

今回の実験では，デコード時のメモリ制限の関係上，提案手法と高次 *n*-gram の言語モデルとを併用した実験は実施できていない．今後，このような追加実験を実施し，言語モデルの次数の影響を取り除いた上で，提案手法の有効性を分析する必要がある．

文献

[1] M. Nagao. A framework of a mechanical translation between Japanese and English by analogy principle. In the International NATO Symposium on Artificial and Human Intelligence, 1981.

[2]P. F. Brown, S. A. D. Pietra, V. J. D. Pietra, and R. L. Mercer. The mathematics of statistical machine translation: Parameter estimation. In Computational Linguistics,. pages 19(2):263–311, 1993

[3] A. Fujii, M. Utiyama, M. Yamamoto, and T. Utsuro. Overview of the Patent Translation Task at the NTCIR-7Workshop. In Proc. of NTCIR-7,pp.389-400, 2008.

[4]A. Stolcke. SRILM - An Extensible Language Modeling Toolkit. In Proceedings of International Conference on Spoken Language Processing, 2002.

[5] P. 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. Moses: Open Source Toolkit for Statistical Machine Translation. Proc. of ACL, demonstration session, 2007.

[6]K. Papineni, S. Roukos, T.Ward, andW.-J. Zhu. Bleu: a method for automatic evaluation of machine translation.In Proc. of ACL, pp. 311–318, 2002.