自然言語処理　144－1
(2001. 7. 16)

対訳文書から自動抽出した用語対訳による機械翻訳の訳語精度向上

出羽達也
(株)東芝 研究開発センター

〒212-8582 川崎市幸区小向東芝町 1
tel：044-549-2239
e-mail：tatsuya.izuha@toshiba.co.jp

あらまし　対訳文書から自動抽出した用語対訳による機械翻訳の訳語改善効果を，特許抄録 20 件を評価データとして検証した．評価データと同分野の対訳文書 2,000 組から自動抽出した用語対訳を利用して機械翻訳を行うことにより，全単名詞句の 23%で訳語の改善が見られた．これは既存の専門用語辞書(20%)を上回る効果である．また，類似文書検索機能を用いて得られた少数(10 件)の対訳文書からでも専門用語辞書と同等の訳語改善効果が得られることを確認した．これは，ユーザが用語対訳を抽出するための対訳文書を予め分野別に分類しておく必要がないということを意味する．

キーワード　機械翻訳，対訳文書，用語対訳

Machine Translation Using Bilingual Term Entries Extracted from Parallel Texts

Tatsuya Izuha
R&D Center, Toshiba Corporation

1, Komukai Toshiba-cho, Saiwai-ku, Kawasaki, 212-8582
tel：044-549-2239
e-mail：tatsuya.izuha@toshiba.co.jp

Abstract　Patent summaries are machine-translated using bilingual term entries extracted from parallel texts for evaluation. The results show that bilingual term entries extracted from 2,000 pairs of parallel texts introduce more improvements than the existing technical term dictionary. The result also shows that even with fewer pairs of parallel texts found by similar text retrieval, substantially good translation is obtained, suggesting that parallel texts to be used do not need to be classified into fields prior to term extraction.

key words　machine translation, parallel text, bilingual term entry

1．はじめに

技術文書の高精度な機械翻訳には，その分野の専門用語辞書が不可欠である．機械翻訳システム用の専門用語辞書がいくつか開発されているが，多くは特定のシステム専用のもので，「医学」「ビジネス」といった比較的広い範囲をカバーすることが多い．そのため，専門性の高い技術文書を翻訳しようとすると，(1)語彙が十分に網羅されていない，(2)訳語を一つに絞り込めない，等の問題が生じる．もっと狭い分野毎に専門用語辞書が整備されていることが望ましいが，専門分野の数は非常に多く，しかも新しい分野が次々と生まれていることから，辞書の構築・維持には多大なコストがかかる．

上述のような事情を背景に，対訳コーパスから対訳辞書を(半)自動構築するための手法が近年盛んに提案されている[1][2][3][4][5]．しかしそれらの手法も，機械翻訳システムで実際に利用するには当該分野の専門家が出力をチェックする必要があり，問題の根本的な解決には至っていない．そこで本稿では，対訳文書から自動抽出した用語対訳を人手を介さずに機械翻訳システムで直接利用することにより，翻訳精度の向上にどの程度寄与することができるか検証することにした．

本稿は以下のように構成される．次節では，本稿における試み全体の概要を説明する．3～5 節では，各コンポーネントについて詳細に説明する．6 節では，特許抄録を用いた評価結果を示し，7 節でその結果について考察する．最後に 8 節の結論をもって本稿を締めくくる．

2. 概要

対訳文書から自動抽出した用語対訳を用いた機械翻訳の処理の流れを，日英翻訳の場合について示したのが図 1 である．

入力として日本語文書が与えられると，これと内容の類似した対訳文書を対訳文書データベースから検索する．その際，内容の類似性は原言語(日本語)側の類似性のみに基づいて判定する．検索された対訳文書から用語対訳の候補を自動抽出し，その中から実際に翻訳で利用するものをある基準で選択する．選択された用語対訳は，一時的にユーザ辞書に登録しておく．入力文書を機械翻訳する際に，システム辞書（日英対訳辞書）に加えて，このようなユーザ辞書を用いることにより，対訳文書から抽出した用語対訳が翻訳結果に反映される．

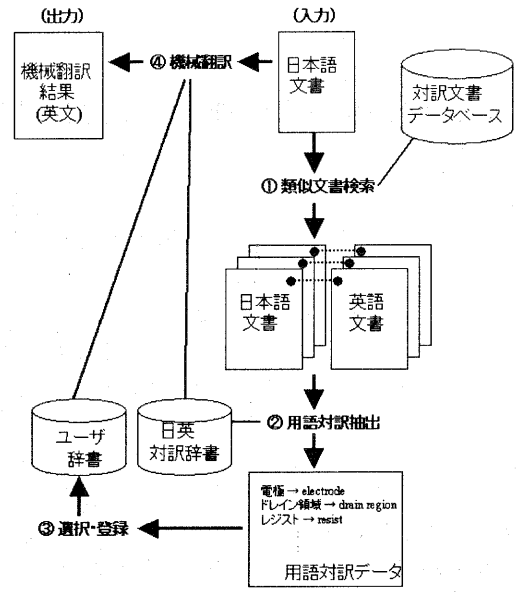

図 1 ：対訳文書から抽出した用語対訳を利用した日英機械翻訳処理の流れ

3. 類似文書検索

入力文書と分野の異なる対訳文書から用語対訳を抽出しても訳語精度の向上は期待できない．しかし，対訳文書群を予め細かく分類しておかなければならないというのではユーザの負担が大きい．そこで本稿では，類似文書検索コンポーネントを用意し，様々な分野の文書を未分類のまま格納した対訳文書データベースの中から入力文書と内容の類似したものを取り出し，そこから用語対訳を抽出できるようにした．これにより，より効果的な用語対訳が抽出できると期待できるが，反面，一つの文書を翻訳しようとする度に用語対訳抽出処理を行わなければならなくなるため，比較的少量の対訳文書から抽出しなければならないという制約が生じる．

本稿で用意した類似文書検索コンポーネントはベクトル空間モデル[6]に基づくもので，各対訳文書をその原言語側の内容を表す特徴ベクトルで表現する．ベクトルの各次元は文書中に出現する各単語に対応しており，その値は TF-IDF 値で重み付けされている．文書間の類似度は特徴ベクトル間

の cosine 値として計算される．入力文書が与えられると，データベース中の各文書との間の類似度を計算し，類似度の降順にソートされた対訳文書のリストを出力する．

対訳文書データベースが予め十分詳細に分類されているならば，類似文書検索コンポーネントは必要ない．例えば，特許明細書にはすべて「国際特許分類(IPC)」[7]コードが付与されている．IPC はセクション，クラス，サブクラス，メイングループ，サブグループから成る階層分類で，現時点で最新の第 7 版では，サブグループの数は約 69,000 におよぶ．

4. 対訳文書からの用語対訳抽出

本稿では，熊野らの方法[1]を用いて対訳文書から用語対訳を抽出する．この手法は，統計情報と言語情報を併用することを特徴としており，比較的少量の対訳文書からでも精度良く用語対訳が抽出できるものと思われる．また，厳密な 1 対 1 の文対応を前提としていないので，実際に世の中に存在する対訳文書を処理する上で現実的である．図 2 を参照しながら，おおまかな処理の流れを説明する．

(1) 日本語文書，英語文書をそれぞれ文単位に分割する．
(2) 日英対訳辞書を参照しながら各日本語文に対応する英語文に推定する．対訳辞書には，日英機械翻訳システム用の辞書を用いる．
(3) 各日本語文から用語(JW)を抽出する．
(4) JW が出現する日本語文に対応する英語文から単語 n-gram を生成し，JW の訳語候補(EW_i)とする．
(5) JW に対する EW_i の対訳確信度 TL(Translation Likelihood)を計算する．

ステップ(3)では，用語として単独名詞，複合名詞，未知語を抽出する．熊野らは専門用語辞書構築という目的から，単独名詞は対象外としているが，本稿では分野に応じた訳し分けという観点から単独名詞も含めることにした．

ステップ(5)では，言語情報に基づいた対訳確信度 TLL(Translation Likelihood based on Linguistic information)と，統計情報に基づく対訳確信度 TLS(Translation Likelihood based on Statistical information)をまず求め，TL は TLL と TLS の加重平均で算出する．

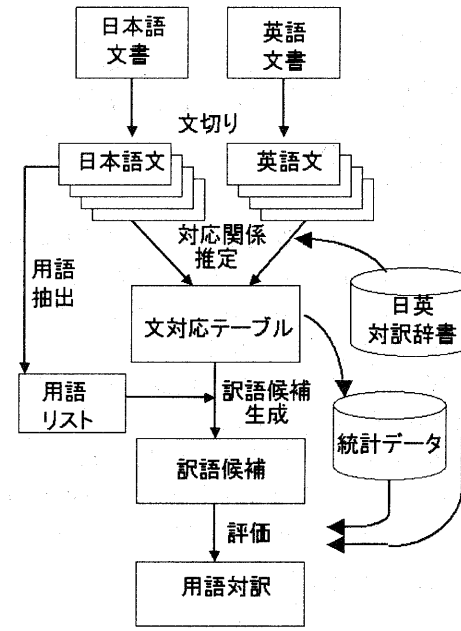

図 2 ：対訳文書からの用語対訳抽出処理の流れ

訳語候補 EW_i の TLL の値は次の 2 つの仮説に基づいて求める．

(仮説 1) EW_i を構成する要素単語の数と JW を構成する要素単語の数が近いほど，JW に対する EW_i の対訳確信度は大きい．

(仮説 2) EW_i の要素単語の中で，JW の要素単語のどれかと対訳関係*にあるものの割合が大きいほど，JW に対する EW_i の対訳確信度は大きい．

(仮説 2)における対訳関係の判定に用いる日英対訳辞書としては，文対応の場合と同様に，日英機械翻訳システム用の辞書を用いる．

JW, EW_i の要素単語の数がそれぞれ k, l で，そのうち対訳関係にある要素単語の数が x であると

* 日英対訳辞書で与えられる日本語単語 jw の英訳語の中に英単語 ew が含まれているとき，ew は jw と対訳関係にあるという．

き，JW に対する EW_i の TLL(JW, EW_i)は次式で与えられる．

$$TLL(JW, EW_i) = \frac{P \times \min(k, l) + \alpha P \times x}{P \times k + \alpha P \times k}$$

$$= \frac{\min(k, l) + \alpha x}{(1+\alpha)k}$$

分子第 1 項では，(仮説 1)に基づき要素単語数の一致に応じて単位スコア P が与えられる．第 2 項では，対訳関係にある要素単語に対して P の α 倍のスコアが与えられる．本稿では α=2 とした．分母は，JW に対して(仮説 1)と(仮説 2)を最大限満たす仮想訳語のスコアである．

一方，JW が m 個の文に出現し，対応する英文のうち n 個に EW_i が出現するとき，JW に対する EW_i の TLS(JW, EW_i)は次式で与えられる．

$$TSL(JW, EW_i) = \frac{n}{m}$$

5. 翻訳で利用する用語対訳の選択

対訳文書から抽出された用語対訳は，ユーザ辞書に一時的に登録することにより翻訳結果に反映される．前節で説明した処理の出力として，一つの日本語用語 JW に対して複数の訳語候補 EW_i が対訳確信度 TL とともに得られる．これに対して何らかの基準に基づき，複数の EW_i の中から一つを選択して JW の訳語としてユーザ辞書に登録するか，全く何も登録しないかを判定する必要がある．最も単純な基準は，TL に閾値を設定し，その閾値を超えた EW_i の中から TL が最大のものを選択するというものである．しかしその場合，TL の最大値と 2 番目に大きい値との差が小さいときに不適当な用語対訳を選択してしまう可能性が大きくなる．そこで本稿では，以下の基準を満たすとき，TL が最大の EW_i を JW の訳語としてユーザ辞書に登録することにした．

(条件1) $\cap$ ((条件2－1) $\cup$ (条件2－2))

(条件 1)

$$TL \geq \beta$$

(条件 2-1)

$$TLL > (TL\text{の値が}2\text{番目に大きい}EW_i\text{の}TLL)$$

(条件 2-2)

$$\ln \frac{n_1}{n_2} - Z_{0.9} \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \geq \theta$$

(条件 1)において，TL の閾値 β の値は予備実験の結果から 0.5 に設定した．(条件 2-1)では，TLL の値は比較的信頼できることから，僅差であっても TLL の値がより大きい訳語候補を無条件に採用することにした．(条件 2-2)は Dagan ら[8]によって提案された基準である．ここで，n_1 は TL 最大の訳語候補が出現した文の数，n_2 は TL が 2 番目に大きい訳語候補が出現した文の数，$Z_{0.9}$(=1.282)は 90％信頼係数，θ は実験的に定める定数で，本稿では θ=0.2 とした．

訳語として採用するか否かの決定が(条件 2-2)に依存するケースを図 3 に示す．

図 3 の第一の例では，日本語用語「電界/強度」('/'は要素単語の区切りを表す)に対する訳語候補として"field strength"と"field intensity"の 2 つが抽出されている．TL が最も大きい"field strength"は，TL=0.67 であるから(条件 1)を満たしているが，TL が 2 番目に大きい"field intensity"と TLL の値が同じ(=0.67)であるため(条件 2-1)を満たしていない．さらに，以下に示すように(条件 2-2)も満たしていないため，この例からはユーザ辞書には何も登録されない．

【電界/強度】 [m=6]

[1] "field strength"
TL=0.67 (*TLL*=0.67 [x=1], *TLS*=0.67 [n=4])
[2] "field intensity"
TL=0.58 (*TLL*=0.67 [x=1], *TLS*=0.33 [n=2])

【酸化/膜】 [m=26]

[1] "oxide film"
TL=0.76 (*TLL*=0.67 [x=1], *TLS* =0.96 [n=25])
[2] "nitride film"
TL=0.60 (*TLL* =0.67 [x=1], *TLS* =0.46 [n=12])

図 3：対訳文書から抽出された用語対訳の例

$$\ln\frac{n_1}{n_2} - Z_{0.9}\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}$$

$$= \ln\frac{4}{2} - 1.282\sqrt{\frac{1}{4}+\frac{1}{2}} = -0.268 < \theta(=0.2)$$

図3の第二の例では，日本語用語「酸化/膜」に対する訳語候補として"oxide film"と"nitride film"の2つが抽出されている．第一の例と同様に(条件1)を満たし，(条件2-2)は満たしていない．しかし，以下に示すように(条件2-2)を満たしているので，「酸化膜」と"oxide film"の組がユーザ辞書に登録される．

$$\ln\frac{25}{12} - 1.282\sqrt{\frac{1}{25}+\frac{1}{12}} = 0.383 > \theta(=0.2)$$

6. 評価

前節までに述べた方法で，対訳文書から抽出した用語対訳を利用して日本語の特許抄録20件を機械翻訳システムで英語に翻訳し，人手による翻訳結果と比較することにより，本手法の評価を行った．対訳文書データベースには，日英対訳特許抄録(日本語で書かれた特許抄録と，それを人手で英訳したもの)2,000組を用いた．評価用のテキスト，データベース中のテキストともにすべてIPCのメイングループH01L 21/00 (半導体装置または固体装置またはそれらの部品の製造または処理に適用される方法または装置) に属する．また，本節の評価で使用した日英機械翻訳システムは25万語のシステム辞書を備えている．

対訳文書から抽出した用語対訳を利用せずに機械翻訳にかけた結果を評価のベースラインとし，人手で翻訳した結果を参照しながら訳語の改善・悪化箇所をカウントした．機械翻訳の出力が人手翻訳結果とは異なるが正しい翻訳結果と思われる箇所もいくつかあったが，評価を容易にするため，人手翻訳結果に近付けば改善，遠ざかれば悪化としてカウントした．

評価の対象は単名詞句(base NP)[9]の訳語に限定している．また，改善・悪化のカウントも単名詞句を単位に行った．例えば，複数の要素単語から構成される複合名詞において，一つの要素単語だけで訳語が改善された場合も，二つ以上の要素単語で訳語が改善された場合も同じように1箇所の改善としてカウントされる．評価に用いた20件の特許抄録のベースライン翻訳結果には，延べ666個の単名詞句が出現する．

日本語用語	旧訳語	新訳語
ゲート電極 領域	gate 電極 domain	gate electrode region
ソース オーミック コンタクト層	sauce オーミック contact layer	source ohmic contact layer
電子線 レジスト	electronic line register strike	electron beam resist
窒化水素	nitriding hydrogen	hydrogen nitride
最上層	best layer	uppermost layer

図4：対訳文書から抽出した用語対訳による訳語改善例

評価結果を表1に示す．「上位10文書」の列と「上位100文書」の列はそれぞれ，類似文書検索で得られた上位10件および上位100件の対訳文書から用語対訳を抽出した場合の改善箇所と悪化箇所の数を示している．「全文書」の列は，類似文書検索による絞込みを行わずに，データベース中のすべての対訳文書2,000組から用語対訳を抽出した場合の結果である．「専門用語辞書」の列は，対訳文書から抽出した用語対訳の代わりに既存の専門用語辞書を用いた場合の訳語改善効果である．専門用語辞書には様々な分野のものがあるが*，ここでは半導体装置に最も関連があると思われる「電気・電子」を利用した．語彙数は約3万8千語である．「専門用語＋全文書」の列は，上述の専門用語辞書と全対訳文書2,000組から抽出した用語対訳の両方を用いた場合の訳語改善効果である．

図4は，対訳文書から抽出した用語対訳によって訳語が改善された例である．

7. 考察

入力文書と同じ分野に属する対訳文書であれば，その量が多いほど，そこからより多くの有用な用語対訳を高い精度で抽出できる．その意味で「全

* 「ビジネス」「化学・金属」「生物」「物理・数学」「情報・コンピュータ・通信」「機械・プラント・建築」「電気・電子」「医学」の8種類

当然のことながら，(仮説 1)は常に成り立つわけではない．TLL(言語情報に基づく対訳確信度)の算出方法あるいは(条件 2-1)を詳細化することにより，ある程度対処は可能と思われる．

現在の実装では，前置詞で始まる単語 n-gram は訳語候補から除外している．そのため，"OFF current"のような訳語候補が生成されない．このような誤りは実装の簡単な修正で防止できる．

(仮説 1)(仮説 2)とも日本語用語の要素単語の区切りが正しいことを前提としているため，単語区切りを誤ると全く誤った訳語推定を行ってしまう．単語区切りの誤りがある程度生じるのは避けられないが，頻度の低い訳語候補を棄却することにより，誤った用語対訳の抽出をかなり防ぐことができると思われる．正しい訳語候補まで棄却される可能性もあるが，今後の実験・評価を通じてバランス点を見極める必要がある．

8. 結論

対訳文書から自動抽出した用語対訳による機械翻訳の訳語改善効果を，特許抄録 20 件を評価データとして検証した．評価データと同分野の対訳文書 2,000 組から自動抽出した用語対訳を利用して機械翻訳を行うことにより，全単名詞句の 23%で訳語の改善が見られた．これは既存の専門用語辞書(20%)を上回る効果である．また，類似文書検索機能を用いて得られた少数(10 件)の対訳文書からでも専門用語辞書と同等の訳語改善効果が得られることを確認した．これは，ユーザが用語対訳を抽出するための対訳文書を予め分野別に分類しておく必要がないということを意味する．

今後は異なる分野のテキストでも評価を行いながら，副作用の要因の洗い出しと，改善効果の拡大を図る．

参考文献

[1] 熊野明, 平川秀樹. "対訳文書からの機械翻訳専門用語辞書作成", 情報処理学会論文誌, Vol.35, No.11, pp.2283-2290, 1994.

[2] M.Haruno, S.Ikehara, T.Yamazaki. "Learning Bilingual Collocations by Word-level Sorting", COLING 96, pp.525-530, 1996.

[3] 北村美穂子, 松本裕治. "対訳コーパスを利用した対訳表現の自動抽出", 情報処理学会論文誌, Vol.38, No.4, pp.727-736, 1997.

[4] F.Smadja, K.R.McKeown, V.Hatzivassiloglou. "Translation Collocations for Bilingual Lexicons: A Statistical Approach", Computational Linguistics, 22, 1, pp.1-38, 1996

[5] I.D.Melamed, "Automatic Construction of Clean Broad-coverage Translation Lexicons", Proceedings of 2nd Conference of Association for Machine Translation in the Americas, pp.125-134, 1996

[6] G.Salton, A.Wong, C.Yang, "A Vector Space Model for Information Retrieval", Communications of the ACM, 18, 11, pp.613-620, 1975

[7] "International Patent Classification (IPC)", http://www.wipo.int/classifications/en/index.html

[8] I.Dagan, A.Itai, "Word Sense Disambiguation Using a Second Language Monolingual Corpus", Computational Linguistics, 20, 4, pp.563-596, 1994.

[9] L.A.Ramshaw, M.P.Marcus, "Text Chunking Using Transformation-based Learning", Proceedings of the 3rd Workshop on Very Large Corpora, pp.88-94, 1995.