対訳文書からの専門用語辞書作成

7Q－5

熊野 明　平川秀樹
{kmn,hirakawa}@isl.rdc.toshiba.co.jp
(株)東芝 研究開発センター

1. はじめに

対訳文書データ(対訳コーパス)は自然言語処理における情報の源として注目されており、各種の知識獲得に関する研究が行われている。日英対訳コーパスからn-gramを作成して対訳辞書を半自動生成する研究[1]や、英語とフランス語の間で名詞句の対応を抽出する研究[2]などがあるが、これらは、主に統計情報に基いたものである。一方、言語情報に基いたものとして、機械翻訳辞書を利用して英日対訳コーパスから専門用語対訳辞書を自動作成する研究[3]があるが、機械翻訳辞書から生成可能な訳語以外は抽出することができない。

本稿では、既存の対訳文書から合成語と未知語を専門用語として抽出し、訳語を推定することによって対訳辞書を作成する方式について述べる。

2. 辞書作成のアプローチ

本研究は、対訳文書から専門用語辞書を自動作成することを目的としている。一般の日英対訳文書では、言語間の文単位の対応が単純でない場合が多く、特に特許明細書などでは、文書構成の違いにより記述順序も大きく異なる。我々はこのような対訳文書に対してもロバストな方式を開発するため、言語的な情報と統計的な情報を統合して利用する。言語的な情報には辞書的・構文的・意味的知識が含まれているので、文書の断片からでも語句の対応を判断できるという特徴があり、統計的な情報には多くの実例から抽象化した知識が含まれているので、雑音に強いという特徴がある。両者の統合することにより、機械翻訳対訳辞書などの言語情報利用だけでは対訳抽出の困難な未知語も処理が可能になる。

本方式では、以下の手順で専門用語辞書を作成する。この処理の流れを図1に示す。

[1] 文の抽出
　日本語文書と英語文書から文を抽出する。
[2] 文間の対応関係の推定
　日本文と英文の対応関係を推定する。
[3] 専門用語の抽出
　日本文から専門用語の候補を抽出する。
[4] 用語の英訳語候補の生成
　対応英文から、専門用語の訳語候補を生成する。
[5] 用語の英訳語候補の評価
　訳語候補を評価して最も確かなものを選定する。

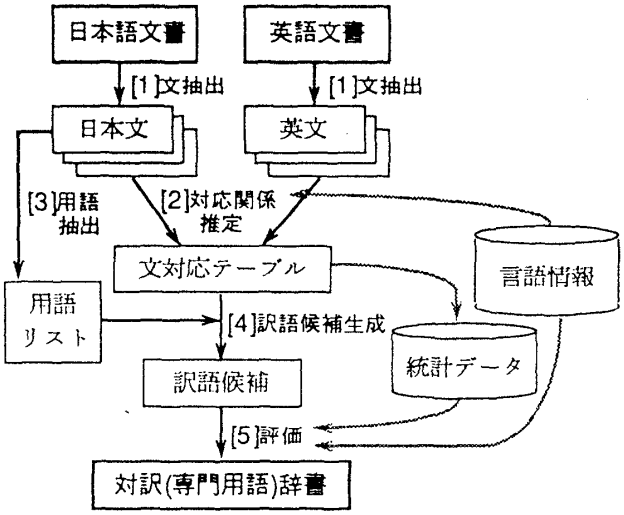

図1: 辞書作成処理の流れ

3. 対応関係の推定

対象とする対訳文書の文対応関係の推定には、文書中の記述順序を利用できないため、機械翻訳の日英対訳辞書のもつ訳語情報を利用した。日本文中の各内容語に対して機械翻訳対訳辞書を参照し、その訳語候補と英文中の内容語との対応関係を文間の対応関係とした。

ある日本文中の内容語 m語のうち、訳語候補が英文中の内容語と一致するものがx語、英文中の内容語 n語のうち日本語内容語のいずれかの訳語候補と一致するものがy語の場合、対応関係確信度を $\frac{x+y}{m+n}$ で計算する。各日本文に対して、対応関係確信度が大きい英文から順に対応文として推定し、文対応テーブルに格納する。

4. 用語の抽出と訳語候補の生成

日本語文書中から抽出する専門用語の候補として、次の2種類を扱った。

(A) 名詞連続の合成語 (サ変動詞を含む)
　[例]「オープンビット線方式」、「最密充填」
(B) 未知語 (名詞およびサ変動詞)
　[例]「積層する」、「ポリッシング」

日本文を構文解析してこれらの用語を抽出し、用語リストを作成する。

各用語の訳語候補生成には、n-gramデータを利用する。ある抽出用語の日本語文書における文出現頻度がN、英語文書の全英文数がMの場合、$N\times M$の対応関係確信度のうち上位N個の値をもつ英文から、n-gramデータを抽出する。この結果、英文中の任意の単語列を訳語

Building a Technical Term Dictionary from Parallel Texts
Akira KUMANO and Hideki HIRAKAWA
Research and Development Center, Toshiba Corp.
1, Komukai Toshiba-cho, Saiwai-ku, Kawasaki, 210, Japan

候補とすることができる。

抽出対象英文全体から作成したn-gramデータ(単語列)のうち、英文における出現頻度の高いものから順に訳語候補 $Ec_1, Ec_2, \ldots Ec_j$ とする。今回専門用語として抽出した合成語や未知語は名詞であることから、動詞 be を含む単語列、前置詞・冠詞で始まるか終わる単語列は予め訳語候補から除外した。

5. 訳語候補の評価

訳語候補 Ec_i が用語 Jw の訳語である確信度を、2つの確信度の関数として定義する。

$$TL(Jw, Ec_i) = F(TLS(Jw, Ec_i), TLL(Jw, Ec_i))$$

5.1 統計情報の利用

$TLS(Jw, Ec_i)$は統計情報に基いた対訳確信度である。言語間の語句の対応関係を統計処理で推定する方法には文献[2]で利用している方法等があるが、文単位の対応関係を前提としており、今回の対象文書に適しているかは検討が必要である。ここでは、単純に出現文頻度を利用し、訳語候補が対応候補文に現れる確率で与える。

$$TLS(Jw, Ec_i) = \frac{Ec_i\text{が現れる英文数}}{Jw\text{が現れる日本文数}}$$

5.2 言語情報の利用

$TLL(Jw, Ec_i)$は言語情報に基いた対訳類似スコアである。ここでは次の仮説を利用する。

[仮説]

(a) 用語 Jwと訳語候補 Ec_iとは構成要素単語数が近いほど対応が確からしい。

(b) 用語 Jw中の各語と訳語候補 Ec_i中の各語に対訳関係が多いほど対応が確からしい。

いま JwとEc_iを次のように構成語の集合とみなす。

$$Jw = \{wj_1, wj_2, \ldots wj_k\},\ Ec_i = \{we_1, we_2, \ldots we_l\}$$

仮説により、すべてのweがいずれかのwjと過不足なく対訳関係のある訳語候補 Ec_iを、最も確かな仮想訳語と仮定する。対訳類似スコアの計算は、仮想訳語との比較で行う。Ec_iの構成語 weのうちいずれかのwjと対訳関係のあるものには得点3P、対訳関係はないが語数の対応がとれるものに得点Pを与える。$TLL(Jw, Ec_i)$ は、Ec_iの得点と仮想訳語の得点($=k\times3P$)との比で与える。

[例]「オーブン/ビット/線/方式」(k=4)

open bit line:	(3×3P)/12P = 0.75
bit line configuration:	(2×3P+P)/12P = 0.58
open bit line configuration:	(3×3P+P)/12P = 0.83

5.3 統計情報と言語情報の統合

用語Jwに対する訳語候補Ec_iの確信度を、次の式で再定義する。

$$TL(Jw, Ec_i) = \frac{pTLS(Jw,Ec_i) + qTLL(Jw,Ec_i)}{p+q}$$

pとqの比を一定にした予備実験の結果、文出現頻度Nの小さい場合に $TLS(Jw, Ec_i)$ の低い値が大きく影響し、正しい Ec_iに対して対訳確信度を低くすることがわかった。このため、pとqの比をNの関数として与えた。

6. 実験

同一分野に関する日本語の特許明細書7件とそれを技術者が翻訳した英訳文を使って実験を行った。日本語文書のサイズは、全体で 2,148文、99,286文字(平均 307文、14,184文字)である。

機械翻訳辞書作成経験者が別途選定した正解に対して、本方式で推定した訳語(第1位と上位3位)が一致する率を表1に示す。統計情報が推定精度に及ぼす影響を調べるために、対象文書量を変えて処理を行った。上段は1文書ごとの処理結果の平均、下段は7文書を統合して処理した結果である。いずれも統合実験の結果が平均を上回っているが、特に未知語の正答率が上昇している。未知語に対する訳語推定処理は、統計情報を利用している部分が大きいため、文書量を増やすことによる効果が現れていることがわかる。

表1: 推定訳語の正答率

処理単位	合成語 (頻度 3,224)		未知語 (頻度 389)	
	第1訳語	上位3訳語	第1訳語	上位3訳語
1文書	71.7% (2,312)	82.5% (2,661)	30.1% (117)	52.4% (204)
7文書	72.9% (2,349)	83.3% (2,680)	54.0% (210)	65.0% (253)

7. 結論

言語情報と統計情報を利用することで、対訳文書から合成語と未知語の両者に対して訳語を推定することができた。合成語では約70%の精度が得られ、未知語に対する精度は相対的に低いが、対象文書量を増やすことで改良できる見通しを得た。

今後は、他の統計情報を用いた手法を検討するとともに、より深い言語情報を利用することで、訳語推定の精度を向上させる予定である。

参考文献

[1] 野美山浩: コーパスからの対訳辞書の半自動生成, 情報処理学会第47回全国大会 6P-8 (1993)

[2] Julian Kupiec: An Algorithm for Finding Noun Phrase Correspondences in Bilingual Corpora, In *Proc. of the 31st Annual Meeting of the ACL*, pp. 17-22 (1993)

[3] 山本由紀雄, 坂本仁: 対訳コーパスを用いた専門用語対訳辞書の作成, 情報処理学会研究報告, NL94-12 (1993)