対訳コーパスからの統計的訳語対抽出における抽出不能語の調査

辻　慶太
国立情報学研究所（keita@nii.ac.jp）

近年，対訳コーパスの入手可能性が高まるにつれ，そこから訳語対を自動抽出する研究が盛んに行われている。一般に，対訳コーパスが存在する状況・分野には，英和辞書や学術用語集のような，二言語辞書・対訳用語リストが存在すると考えるのが自然である。そうした辞書・用語リストに載っている訳語対を，対訳コーパスからあらためて抽出してもメリットは少ない。対訳コーパスから自動抽出する価値のある訳語対は，少なくとも辞書・用語リストに載っていない訳語対である。

コーパス中の頻度が高い訳語対は，一般的あるいは分野の中心的な訳語対であり，既に辞書や用語リストに載っている可能性が高い。逆に頻度が低い訳語対は載っていない可能性が高い。以上のような見地から，対訳コーパスからの訳語対自動抽出で重要なのは，低頻度訳語対の抽出であると考える。本研究は，低頻度訳語対は先行研究手法の多くでは抽出が難しいことを示し，新たな方向を提案するものである。

対訳コーパスからの訳語対自動抽出研究では，まず「そこから訳語対を抽出する」範囲をあらかじめ設定する。そのような範囲には，文や段落といった，意味的な対応があると見なされるテキスト部分が，選ばれることが多い。以下，この部分を「セグメント」と呼ぶ。さて，従来の先行研究手法の多くは，同じセグメントに共起する度合いが高い対を，訳語対として抽出している。

だがこの手法には次の致命的な欠陥がある。まず，訳語対 XY があり，語 A が，X の現れるセグメントに必ず現れ，X の現れないセグメントには現れない場合，語 A は訳語対 XY と完全共起していると定義する。さらに A が X と同じ言語に属する場合，語 A は訳語対 XY と内部完全共起していると定義する。また A が X と異なる言語に属する場合は，語 A は訳語対 XY と外部完全共起していると定義する。内部完全共起が起きている場合，従来の語の出現・共起に基づく統計的抽出手法は，Y の訳語が X なのか A なのか決定することが出来ない。また外部完全共起が起きている場合，従来の語の出現・共起に基づく統計的抽出手法は，X の訳語は Y ではなく A であるとしてしまう。即ち，完全共起が起きている場合，訳語対 XY は頻度に基づく統計的手法では適切に抽出できない。

本研究では，国立情報学研究所の学会発表データベースにおける日英著者抄録を対訳コーパス，EDICT を対訳辞書として調査対象にし，以下の点を確認した。

(1)コーパス中の頻度と辞書収載に関する上記の予想は妥当であること。
(2)日英語の一方の頻度が１の場合，非常に高い割合で他語に完全共起されるが，日英語の双方の頻度が２以上になると，完全共起される割合は急激に減ること。
(3)日本語と英語の場合，低頻度の訳語対には，カタカナやローマ字で表記された，借用語系の訳語対が多いこと。

借用語系の訳語対は，翻字パターンに基づいて自動抽出することが可能である。従って，対訳コーパスからの訳語対抽出においては，低頻度訳語対は翻字に基づいて抽出し，残りの訳語対は，従来の頻度に基づく統計的手法で抽出するのが有望な方向と考えられる。