農業関連学術文献の和英タイトルからの同義語抽出

竹﨑あかね　　木浦卓治
農研機構中央農業総合研究センター

要旨

我々は農業分野においてテキストデータから効果的・かつ適切に知識を抽出する基盤となる言語資源開発を行ってきた．本稿では言語資源の整備を効率化するために，学術文献の日本語・英語対訳タイトルから同義関係にある農業用語を抽出する手法を提案する．具体的には，まず，日本語・英語の対訳タイトルから共起度が高い日本語・英語の対訳用語ペアを収集し，既存の言語資源（日本農業シソーラス）の対訳用語ペアに追加する．次に，和英用語をノード，対訳用語ペアをリンクとみなしてネットワークグラフを作成し，リンクによって連結された用語すべてを同義語と定義する．本稿では，提案手法を日本で出版された農業分野における学術文献に適用し，その精度と問題点について論じる．

1　はじめに

情報検索では，文書とクエリ間の同義表現による語彙の不一致を防止するためにシソーラス等を使った検索クエリ拡張が行われ，再現率の向上がはかられる．一方，農業分野において，我々は日本農業シソーラス（JAT）を開発し，学術文献の検索効率化を進めてきた[1]．JAT は，約 57,000 の専門用語が日本語・英語で収録され農業分野の広範な学問領域をカバーしている．しかし，農業分野における学術文献ではしばしば JAT 収録語と同一概念が JAT 未収録の語で表現されており[2]，同義語のさらなる収集が必要と判断された．

農業分野の用語は，生物生産・自然環境・施設などの人工環境・農村社会環境・バイオテクノロジーなど基礎から応用までの広い学問領域を反映した領域固有の用語が多いこと，作物名が地域名や品種名でも表現されること（例：「サツマイモ」は「薩摩芋」・「唐芋」・「琉球芋」・「甘藷」ともいう），他分野と同様に変遷が激しいこと等の特徴がある．農業分野で利用される用語はこれらの特徴により多数かつ多様であり，同義語の人手による網羅的収集は困難である．

同義語を自動獲得するために，農業分野以外ではこれまでに多くの手法が提案されてきた．単一言語コーパスを用い共起度を基準に同義語を自動獲得する手法は効果的であったが，意味的に等価ではない関連語が抽出される傾向があった[3]．そこで，近年では，対訳コーパスを用い，同じ用語に翻訳される用語同士を意味的に等価と仮定して同義語を高精度に抽出する手法が考案されてきた[4-7]．

本稿では，農業分野における学術文献の和英タイトルを対訳コーパスとして利用し，JAT と組み合わせることで和英混在の同義語群を抽出する手法を提案するとともに，提案手法の精度と問題点について論じる．

2　背景

2．1　和英タイトルの対訳コーパスとしての利用

相澤・影浦[7]は学術論文の和英著者キーワードを対訳コーパスとして用いる利点として，研究者が自ら提示するために検索に有用な用語である可能性が高いこと，対象とする文献に固有の用法や最新の話題を反映していることを挙げている．我々は，著者キーワードと同様の利点を持つ，学術文献の和英タイトルを対訳コーパスとして用いることとした．和英タイトルを利用する場合，対訳用語ペアをあらかじめ選定する必要はあるが，要旨や本文と比較して短い文章のためにその選定は容易である．

2．2　共起スコアに基づく対訳コーパスからの対訳関係の選定

対訳コーパスが既に文単位で対応つけられている場合，二言語における単語対が対訳関係である度合いは共起スコアに基づき推定される[8]．これは，対訳コーパスにある対応可能な用語対の共起頻度を求め，その対応のもっともらしさを統計的共起尺度で評価するものである．共起尺度の一つである相互情報量[9]；$I(x,y)$は，下式の通り各単語がそれぞれ独立に出現する回数と対訳文に同時に出現する回数から求める．この式では$p(x)$，$p(y)$が単語x，

単語 y が独立に起こる正規確率を，$p(x,y)$ が単語 x と単語 y の同時確率を示す．

$$I(x,y) = \log_2 \frac{P(x,y)}{P(x)P(y)}$$

なお相互情報量は低頻度の用語で異常に高くなることから，設定した閾値よりも低頻度の用語は対訳用語ペアの候補から削除される[10]．

２．３　同義語群の抽出

対訳コーパスを利用した同義語（言い換え表現）抽出手法では，同じ用語に翻訳される用語同士は意味的に等価と仮定される．これは，例えば，和訳がいずれも“免疫応答”である“Immune response”と“Immunologic reaction”を同義語とみなすものである．これまでに，フレーズ x がフレーズ y に言い換えられる確率を用いた手法[4，5]，ネットワークグラフを用いた手法[6, 7]で同義語の自動獲得が成功している．後者の手法は，和英用語をノード，対訳関係をリンクとみなして大規模な用語グラフを生成するものであり，リンクによって連結された用語をすべて同義語と定義するものである．我々は，和英用語に関わりなく同義語群を抽出するため，ネットワークグラフを用いる手法を採用することとした．

３　方法

３．１　共起スコアに基づく対訳コーパスからの対訳関係の選定

本報では，農林水産研究情報総合センターが構築した，日本農業文献記事索引（JASI：http://www.affrc.go.jp/db_search/jasi）および AGRIS（http://agris.fao.org/）日本版を対象とした．JASI は，1970 年から国内で発行された学術雑誌，国公立農業研究機関・大学の研究報告書等約 500 誌の書誌情報（タイトル・要旨・著者名など）を収集した日本語データベースである．一方，AGRIS 日本版は，JASI と同じ収録範囲から英語の書誌情報を収集した英語データベースである．JASI に収録された書誌データ約 27 万件のうち AGRIS に英語タイトルが収録されていた 58,300 件について，和英タイトルを整理し対訳コーパスとした．日本語コーパス（タイトル）と英語コーパス（タイトル）から各々名詞を抽出し，各コーパスにおける名詞の出現回数を数えた．さらに，対訳コーパスからは和用語（名詞）と英用語（名詞）が同時に出現する回数を数えた．対訳用語ペアの候補となる和英用語対は，下式の相互情報量；MI score を基準に選定した．この式では，P(J)と P(E)が和用語 J，英用語 E が独立に起こる正規確率を，P(J, E)が和用語 J，英用語 E の同時確率を示す．

$$\text{MI score} = \log_2 \frac{P(J,E)}{P(J)P(E)}$$

MI score は出現回数が少ない用語で高く見積もられる．そこで，出現回数が少なすぎる用語はノイズとみなし 5 回以上出現したものを対訳用語ペアの候補とした．対訳用語ペアの選定実験は 2 回行った．実験 1 では和用語の対象を単語・複合語に，英用語を単語に設定した．実験 2 では，和用語・英用語の対象をともに単語・複合語とし，英複合語は JAT に収録があった用語を利用した．選定した和英用語対を確認し，対訳関係にある和英用語対の割合と相互情報量との関係を調査した．また，JAT において和英いずれの表記も収録されていた用語を対訳用語ペアとして利用した．

３．２　ネットワークグラフによる同義語群の抽出

和英用語の関連を可視化するためにオープンソースのバイオインフォマティクス用ソフトウェア Cytoscape（http://www.cytoscape.org/）を利用し（図１），10,871 の対訳用語ペアから和英用語をノード，対訳関係をリンクとする用語グラフを作成した（図２）．用語グラフは非連結の多数のサブグラフで構成され，各々のサブグラフにあるリンクは必ず和用語と英用語を連結し，同一言語の用語間にリンクは存在しない．サブグラフに含まれる用語群を人手により確認し用語間の関連性を調査した．

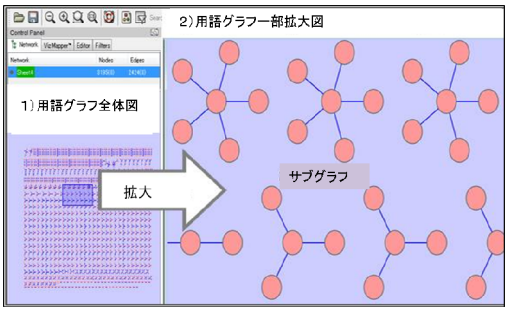

図１　可視化ソフトウェア Cytoscape の画面
用語グラフは非連結の多数のサブグラフで構成される．

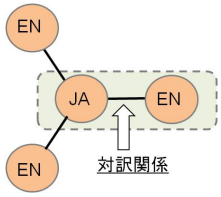

図２　対訳用語ペアから作成した用語グラフ
用語グラフは和用語（JA）・英用語（EN）をノード，対訳関係をリンクとして表される．

４　結果および考察

４．１　共起スコアに基づく対訳コーパスからの対訳関係の選定

対訳関係にある和英用語対の割合は MI score の増加と共に大きくなった．これは対訳コーパス中で共起度が高い和英用語対が対訳関係にあったことを示す．一方，MI score が高い場合でも和英用語対が対訳関係にないエラーの発生が認められた（図３，点線部分）．表 1 は MI score が高かった和英用語対の例を示している．“コナガ（*Plutella　xylostella*）”や“重金属 (heavy metals) ”のように和用語の対訳が複合語の場合，和英用語対は対訳関係になかった．これは実験１で英用語の対象を単語のみにしたため生じたエラーと判断した．そこで，実験２では英用語の対象を単語と複合語にしたところ，[“コナガ”, “*Plutella　xylostella*”] や [“重金属”, “heavy metals”] など対訳関係にある和英用語対の選定に成功した（表２）．しかし，実験２においても複合語による対訳用語ペアの選定エラーを完全に防ぐことはできなかった．これは実験２の英複合語が JAT に収録された用語を利用したものであり，対訳コーパス中の英複合語を網羅していなかったためと考えられる．北村・松本[11]は，各コーパス中に存在する自立語を先頭に複数単語から成る単語対（複合語）を作成し，出現回数の多かった単語列を対訳用語ペアの候補とした．この方法は，既存の言語資源に無い，文献固有の複合語や最新の話題を反映した複合語を対訳用語ペアの候補にできる利点がある．今後は対訳用語ペア候補となる英複合語の抽出手法の開発が課題である．

[“不妊化”, “gamma radiation（ガンマ線）”]のように和用語と英用語の対訳が何らかの関係がある場合にも（“ガンマ線” は“不妊化”を生じさせる原因）対訳関係にある和英用語対として誤って選定された（表２）．この場合，日本語コーパス（タイトル）中の“不妊化”と “ガンマ線”の共起度が高く対訳コーパス中での “不妊化”と “gamma radiation”の共起度も高くなったことによりエラーが生じたと考えられた．誤選定された対訳用語ペアは次の行程でも同義語群を誤って生成する．このため，単一言語コーパスで共起度が高い用語，およびその対訳語が同一の同義語群に含まれる場合はその同義語群を分割するなどの処理が必要と考えられる．

４．２　ネットワークグラフによる同義語群の抽出

用語グラフは多くの種類のサブグラフで構成されていた．このうち 3 用語以上を含むサブグラフは 961 であった．例えば “egg plant”, “*Solanum Melongena*”, “brinjal”, “aubergine” の対訳語はいずれも“ナス”であり，これらは同義語と判定された（図４A）．一方，“hog”, “pig”, “swine”の対訳語は“豚”, “swine”と “*Sus scrofa domesticus*” の対訳語は“ブタ”であったが，“hog”, “pig”, “swine”, “*Sus scrofa domesticus*”はすべて同義語と判定された（図４B）．以上のことから，対訳語を共有する 2 つの用語だけでなくサブグループに含まれる用語は和英混在の同義語群と判断された.

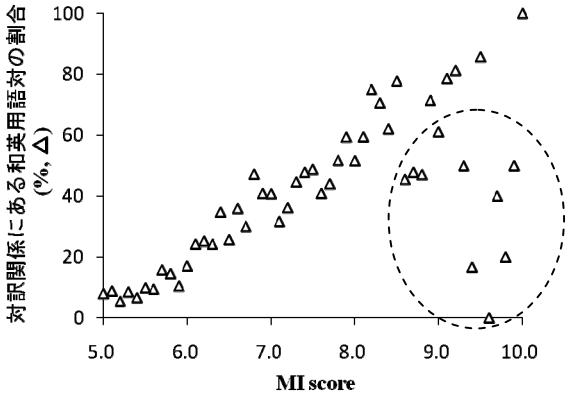

図３　対訳関係にある和英用語対の割合と MI score との関係（実験１）

表１　英単語のみを対象とした場合の高 MI SCORE の和英用語対の例（実験１）

和用語	英用語	MI score	対訳関係 [a]（和用語の対訳）
カゼイン	casein	10.0	対訳
カドミウム	cadmium	9.9	対訳
コナガ	*Plutella*	9.8	非対訳（*Plutella　xylostella*）
重金属	metals	8.9	非対訳（heavy metals）

[a] 和英用語対は人手で対訳関係か否かを確認.

表２　英単語・複合語を対象とした場合の高 MI SCORE の和英用語対の例（実験２）

和用語	英用語	MI score	対訳関係 [a]（和用語の対訳）
アカバネ病	Akabane disease	13.2	対訳
不妊化	gamma radiation	12.8	非対訳（sterilization）
コナガ	*Plutella　xylostella*	9.9	対訳
重金属	heavy metals	8.9	対訳

[a]和英用語対は人手で対訳関係か否かを確認.

本実験では抽出した対訳用語ペアおよび既存のシソーラス；JAT の対訳用語ペアを利用したが，JAT に収録された以外の新たな同義関係を抽出できた．法隆ら[12]は，兄弟関係の語を抽出し，既存シソーラスと組み合わせることで，既存シソーラスに

新たな語の関連性を追加できたとしている．今回抽出した同義語群も JAT と組み合わせることで JAT に新たな同義関係を追加できた．

用語グラフには“cod” (“Pollack”：タラ) ・“COD” (“chemical oxygen demand”： 化学的酸素要求量)のように，複数の語義を持つ多義語が含まれる場合があった．これは語義の曖昧性として自然言語処理では問題となっており，曖昧性解消のために今まで多くの研究が行われてきた．例えば，用語グラフを作成の上グラフ理論を適用して多義語ノード候補を検出する方法[6, 7]，共起関係を利用する方法[13] が報告されている．農業分野では国際連合食糧農業機関（FAO）が近年，用語の多義性による混乱を避けるため用語を概念ベースで整理するシステムを構築し，データの改編作業を開始している[14]．以上のことから今後は，多義語の検出方法とともに，FAO と連動した概念ベースの用語整理方法について検討する必要があると考える．

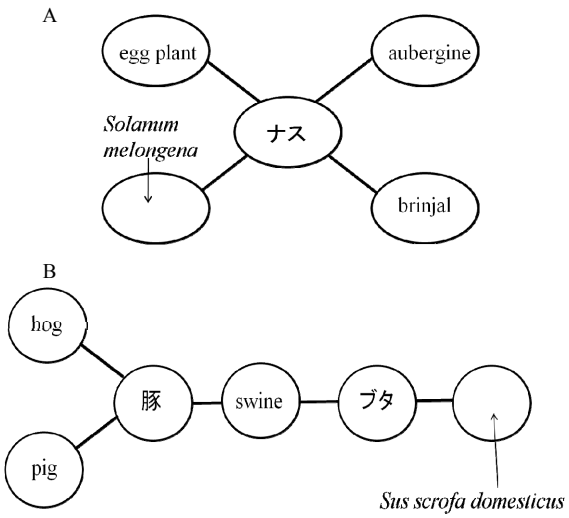

図４　3 用語以上を含むサブグラフの例

５　結論

我々は農業分野の学術文献における和英タイトルから相互情報量を基準として対訳用語ペアを収集した．さらに，和英用語をノード，対訳用語ペアをリンクとみなしてネットワークグラフを作成し同義語群を検出した．3 用語以上を含むサブグラフは 961 あり，これらは和英混在の同義語群と判断した．本手法では対訳用語ペア候補となる英複合語の抽出，単一言語コーパス中の高共起による対訳用語ペアの選定エラーへの対応，多義語の検出が課題として残された．抽出した同義語群は既存シソーラス；JAT と組み合わせることで JAT に新たな同義関係を追加できた．

参考文献

[1] 竹﨑あかね・斉藤三行・岡辺明子（2008）農林水産分野の情報検索に資する言語資源の開発，農業情報研究, 17(1) , pp. 42-49.

[2] 竹﨑あかね・細羽見喬・法隆大輔・木浦卓司（2010）農林水産分野における自動索引付けに有効な言語資源の開発と評価，農業情報研究, 19(1), pp. 10-15.

[3] L. V. D. Plas, I. S. Jacobs, and C. P. Bean（2006） Finding synonyms using automatic word alignment and measure of distributional similarity, Proc. of the COLING/ACL, PP.866-873.

[4] C. Bannard and C. Callison-Burch （ 2005 ） Paraphrasing with bilingual parallel corpora, In Proc. Of ACL ’05 .

[5] 海野裕也・宮尾祐介・辻井潤一（2008）自動獲得された言い換え表現を使った情報検索，言語処理学会第 14 回年次大会，pp.123-126.

[6] N. Kando, A. Aizawa （ 1998 ） Cross-lingual information retrieval using automatically generated multilingual keyword clusters, The 3rd international workshop on information retrieval with asian languages.

[7] 相澤彰子・影浦峡（2000）学術文献の和英著者キーワードを用いた類義語クラスタの自動生成，情報処理学会論文誌, 41(4), pp.1180-1191 .

[8] 言語処理学会編（2009）言語処理学事典，共立出版.

[9] K. W.Church and P. Hanks（1990）Word association norms, mutual information, and lexicography, Computational Linguistics, 16(1), pp.22-29.

[10] P. V. D. Eijk（1993）Automating the acquision of bilingual terminology, EACL’93, pp.113-119.

[11] 北村美穂子・松本裕治（1997）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, 22(1), pp.1-38.

[12] 法隆大輔・木浦卓治・竹崎あかね・斉藤三行・岡辺明子（2009）自動抽出と AGROVOC の組み合わせによる言語資源の整備，農業情報研究, 18(2), pp. 65-71.

[13] D. Yarowsky （ 1995 ） Unsupervised Word Sense Disambiguation Rivaling Supervised Methods, Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistic, Cambridge, MA, pp. 189–196.

[14] M. Sini, S. Rajbhandari, M. Amirhosseini, G. Johannsen, A. Morshed, J. Keizer (2010) The AGROVOC Concept Server Workbench System: Empowering management of agricultural vocabularies with semantics, http://www.fao.org/docrep/012/al055e/al055e00.pdf.