Bilingual KWIC — 対訳抽出の可視化による翻訳支援

小川泰弘　西森 寛敏　外山 勝彦
名古屋大学大学院情報科学研究科
yasuhiro@is.nagoya-u.ac.jp

1　はじめに

機械翻訳システムは，現在，様々な分野で使われているが，翻訳の対象となる文書の内容が専門的である場合，その分野特有の専門用語や定型表現に関する対訳辞書が必要となる．そうした辞書を人手で作成するのはコストが高いため，あらかじめ翻訳された対訳コーパスから専門用語や定型表現の対訳を自動抽出する研究が盛んである [1]．しかし，自動抽出の結果は必ずしも正確ではなく，間違った対訳パターンを抽出したり，対訳パターンの一部だけを抽出する場合がある．また，一つの語に対して複数の訳語を抽出した場合には，訳し分けに関する知見が必要となる．

そこで，対訳パターンを抽出するだけでなく，対訳パターンの各候補を，それが出現した文脈と一緒に表示することによって，ユーザによる対訳パターンの選定を支援し，それにより翻訳作業を支援する手法を提案する．具体的には，対訳抽出の技術と KWIC(KeyWord In Context) 表示を統合し，文単位で対応付けされたパラレル・コーパスから，与えられたキーワードとその対訳パターンの候補をそれぞれ文脈付きで表示する Bilingual KWIC を提案する．

2　Bilingual KWIC の概要

Bilingual KWIC の概観を図 1 に示す．左上のキーワード入力欄にキーワードを入力し，その横の [Search] ボタンを押すと，左側に原言語，右側に対象言語で対応付けられた対訳文を表示する．その際，原言語ではキーワードを中心に，また，対象言語では自動的に推定したその対訳パターンを中心に，それぞれ KWIC 形式で表示する．また，注目する文をマウスでクリックすると，その文全体が下側に表示される．

右上の対訳パターン入力欄には Bilingual KWIC が推定した対訳パターンが表示されるが，それが間違っていたときには，ユーザが自分で別の対訳パターンをここに入力することが可能である．それに加えて，この欄で下向きカーソルキー (↓) を押すと，Bilingual KWIC が推定した対訳パターンの第 2 候補以下が順に表示され，ユーザは別の対訳パターンを選択することが可能である．なお，キーワードおよび対訳パターンともにコーパス中における出現回数がすぐ横に表示され，対訳パターン選定の一助となっている．

また，[Sort] ボタンを用いて出力結果を並び換えることができ，それによって対訳パターンや用例の比較が簡単に行える．図 1 では，対訳語 “floating” に続く語を基準にソートされている．

なお，図 1 では，原言語が日本語，対象言語が英語となっているが，キーワード入力欄に英単語を入力すれば，図 2 のように原言語を英語，対象言語を日本語として対訳パターンの自動抽出が行える．またキーワード，対訳パターンともに複合語を扱うことが可能である．

3　Bilingual KWIC の特徴

Bilingual KWIC は以下のような特徴を持つ．

1. 対訳パターンの抽出における誤りの訂正が容易
2. 派生表現の獲得が容易
3. 訳し分けに関する知見の獲得が容易
4. 他の言語への応用が容易

Bilingual KWIC では，自動対訳抽出における誤りをユーザが簡単に修正できる．図 1 の例では，民法の口語訳とその英語訳[1]から「根抵当権」をキーワードとして検索している．正しい対訳は，“floating mortgage” であるが，自動対訳抽出の結果は “floating” となっている．しかし，図 1 の KWIC 形式で表示された英語コーパスを見れば，“floating mortgage” が正解であることが直観的に理解できる．

[1]今回は 2004 年 12 月に成立した新民法ではなく，文語で書かれた旧民法に対する口語化私案 [2][3] を利用している．英語訳も公式のものは存在しないため，私訳 [4] を利用した．

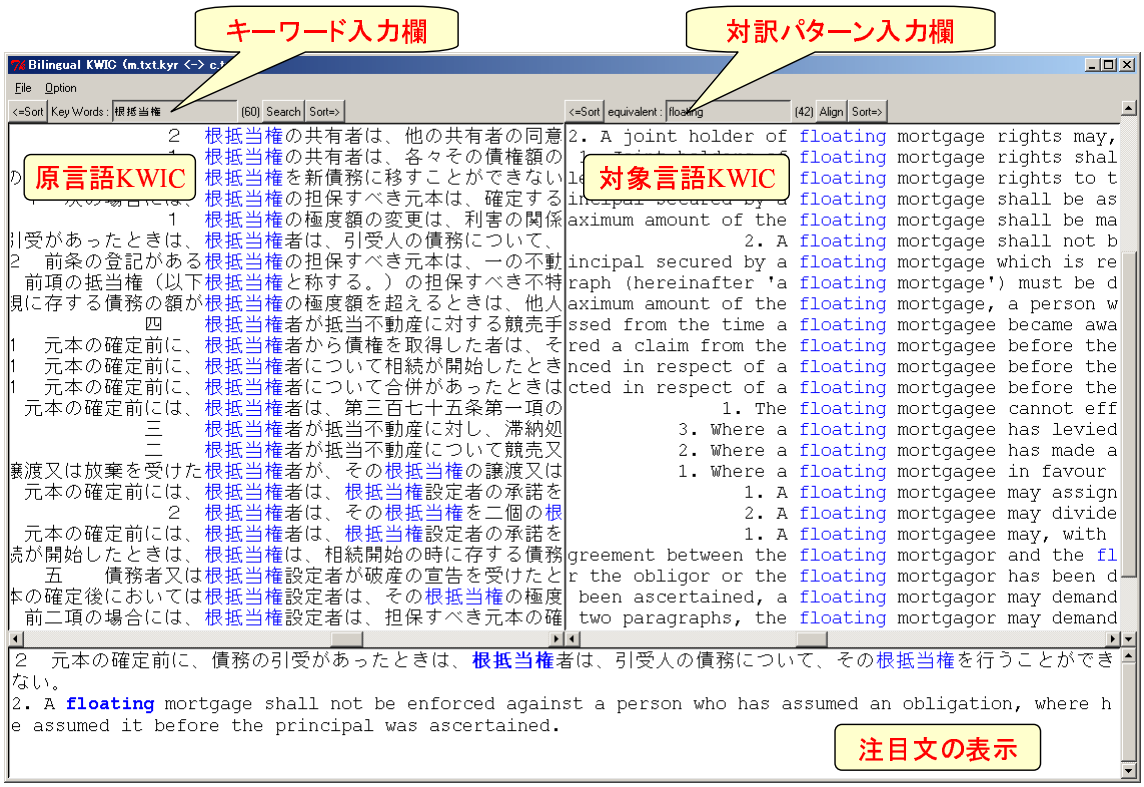

図 1: Bilingual KWIC の概観

さらに，派生表現とその対訳を容易に獲得することができる．図 1 の例では,「根抵当権/floating mortgage」の下に続く用例から「根抵当権者/floating mortgagee」および「根抵当権設定者/floating mortgagor」という関連語とその対訳を得ることができる．

このような特徴は対訳パターンの抽出を支援するときに有用であるが，対訳パターンに加えて，実際の用例も表示されることから，Bilingual KWIC は翻訳支援に対しても有用である．特に訳し分けに関する知見が容易に獲得できる点が優れている．図 2 の例では，“dissolution” の対訳として,「離縁」「解散」「解消」が表示されている．そのような場合には，前後の文脈から，対訳語がどのように使い分けられているかを比較することが容易であり，この例では，養子に関係する場合は「離縁」，法人などの場合は「解散」，婚姻に関しては「解消」という使い分けがなされていることが分かる．

また，現在の Bilingual KWIC は後述するように形態素解析を行わず，文字レベルの情報だけを利用している．よって，他の言語での利用も可能であり，図 3 では，ベトナム憲法 (ベトナム語) とその日本語訳をコーパスとして利用している．

ただし，形態素解析をしないという点から，動詞などの活用する語については，対訳パターンの抽出精度が落ちるという欠点がある．

4 Bilingual KWIC の技術的詳細

4.1 対訳パターンの自動抽出

対訳パターンの自動抽出については，これまでも様々な手法が提案されている．対訳候補のパターン間の類似度の計算についても各種の手法が提案されているが，それらに関しては文献 [1] が詳しい．そうした手法の中から，Bilingual KWIC では以下の Dice 係数を類似度として採用した．

$$Dice(x,y) = \frac{2 \times freq(x,y)}{freq(x) + freq(y)} \qquad (0 \leq Dice(x,y) \leq 1)$$

ここで $freq(x)$ と $freq(y)$ は，入力キーワード x および対訳候補パターン y がそれぞれ原言語コーパスおよび対象言語コーパス中に出現する回数であり，$freq(x,y)$ は，対応付けられた文に x と y が同時に出現する回数である．

なお，先行研究では文献 [5] のように Dice 係数に出現回数の対数値を重みとして掛けたものもあるが，

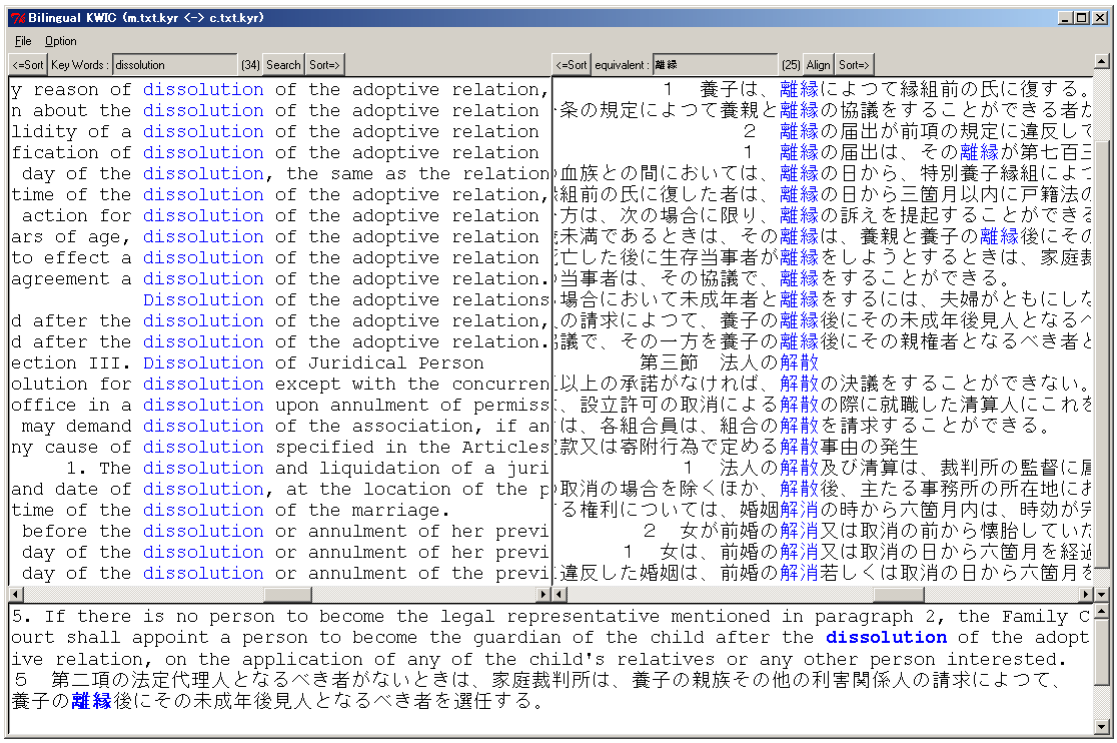

図 2: “dissolution” に対する実行結果

Bilingual KWIC では，キーワードがユーザから与えられ，それに対する対訳パターンを求めることが目的であり，複数のキーワード間での比較は不要であることから，重み付き Dice 係数は採用しなかった．

また Bilingual KWIC では，対訳コーパスから Dice 係数が最大となる対訳パターンをひとつ決定した後，その対訳パターンが含まれていない文を集めて，再度 $freq(x, y)$ を集計し，Dice 係数を計算しなおす．これにより，図 2 のようにキーワードが複数の対訳パターンをもつ場合に，それぞれを抽出することが可能である．

4.2 形態素解析

文献 [5] を含め，先行研究では日本語・英語とも形態素解析するものが多いが，現在の Bilingual KWIC では，形態素解析をせず，文字レベルの情報だけを用いている．ここで文字レベルの情報とは，日本語の平仮名は対訳語に含めない[2]，英語の単語は空白で区切られる，といった情報である．具体的には，日本語は文字 N グラム，英語は単語 N グラムを用いて，ある程度の長さをもつ対訳パターンの候補を求めている．なお，N の最大値は言語ごとに指定可能である．

形態素解析を利用する利点として，対訳パターン抽出の精度向上が期待できる点が挙げられる．特に動詞のように活用する語や，英語の名詞の複数形などは形態素解析をしないと変化形が別の語として認識されてしまう．

その一方で，形態素解析での誤りが対訳抽出に影響を与える，語の一部だけをキーワードとして入力できない，他の言語に応用する場合はその言語に対応した形態素解析システムが必要といった問題もある．Bilingual KWIC では，対訳パターン抽出の誤りを容易に修正できることから，現在は形態素解析を利用していない．しかし，精度向上のため，あらかじめ形態素解析されたコーパスも利用可能とするのは今後の課題である．

なお，現在の Bilingual KWIC では，英単語の変化形などに対処するため，単語の出現回数を数える際に，厳密な一致ではなく接頭語として含まれているものもすべて数え挙げている．例えば，“search” の出現回数を数えるときには “searches”, “searched”, “searcher” なども含めて数えている．これにより，規則変化する語についてある程度対処している．

4.3 Bilingual KWIC の実装

Bilingual KWIC の実装には，Ruby/Tk を用いた．現在は，MS-Windows および Linux 上での動作を確

[2]オプションで含めることも可能．

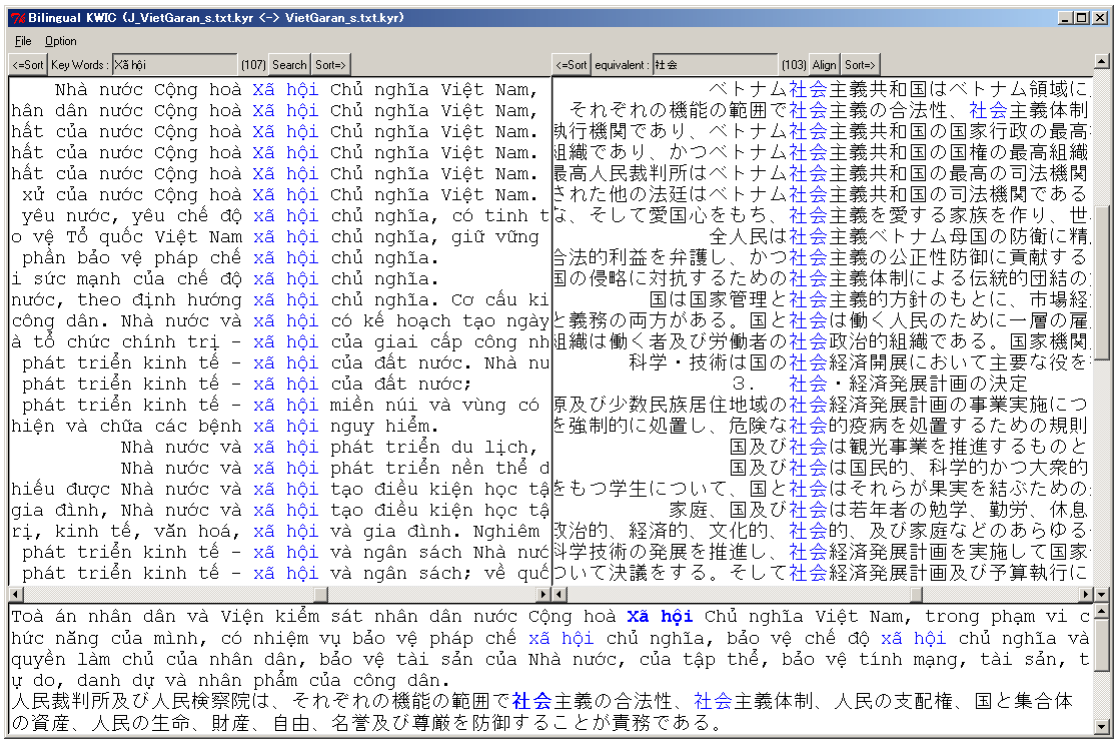

図 3: 日本語–ベトナム語コーパスへの適用

認している．なお，文字列検索の高速化のため，Suffix Array[6] のライブラリである sary[3] を使用している．また，文字コードには Unicode を用いており，図 3 の例のように，各種言語への応用が容易である．

5 まとめ

本稿では，対訳パターン抽出を可視化することで翻訳を支援する Bilingual KWIC を提案した．

先行研究として，文献 [7] が挙げられる．ここに挙げられた翻訳メモリの利用では，与えられたキーワードの日本語コーパスにおける出現を KWIC 形式で表示し，コーパス中においてキーワードと良く共起した日本語単語と英語単語を提示する．ユーザが対訳候補となる英語単語を選ぶと，それに基づいた絞り込み検索を行い，対訳文のペアを別ウィンドウに表示する．ただし，別ウィンドウに表示される対訳文のペアでは，キーワードと対訳候補の部分がそれぞれ下線で明示されるが，中心に揃えて表示される訳ではない．Bilingual KWIC では，最初に対訳候補を自動的に決定する，任意の対訳パターンを指定できる，原言語だけでなく対象言語も同時に KWIC 形式で表示する点が異なる．

今後の課題として，既存の電子辞書や形態素解析を含むヒューリスティックとの組み合わせを可能にすることや，大規模な対訳コーパスにも対応できるよう高速化する点が挙げられる．

なお，Bilingual KWIC は以下のページにおいて公開する予定である．

http://www.kl.i.is.nagoya-u.ac.jp/koyori/

[3] http://namazu.org/ satoru/sary/

参考文献

[1] Mastumoto,Y. and Utsuro, T.: Lexical Knowledge Acquisition, *Handbook of Natural Language Processing*, Dale, R., Moisl, H., and Somers, H.(Eds.), pp.563–610, Marcel Dekker (2000).

[2] 加賀山 茂：民法財産編の口語化草案 (私案)(上), 阪大法学 155 号, pp.185–244 (1990).

[3] 加賀山 茂：民法財産編の口語化草案 (私案)(下), 阪大法学 156 号, pp.495–574 (1990).

[4] Oda, H.: *Basic Japanese Laws*, Oxford University Press (1997).

[5] 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).

[6] 山下 達雄: 用語解説 「Suffix Array」, 人工知能学会誌 Vol. 15, No. 6, p.1142 (2000).

[7] 内山 将夫, 井佐原 均: 日英新聞記事対応付けデータを用いた翻訳メモリと言語横断検索, 情報処理学会第 65 回全国大会第 5 分冊, pp.355–358 (2003).