Wikipedia Infobox の言語間対応付けによる対訳辞書の抽出

胡 寅駿, 林 良彦 (大阪大学言語文化研究科)
永田 昌明 (NTT コミュニケーション科学基礎研究所)

1　はじめに

Wikipedia から構造化された情報･知識を抽出しようとする研究が活発に行われている (例えば [2, 7, 8]). これらの研究には，Wikipedia の持つ多言語性に注目し，いわゆる対訳辞書を抽出しようとする試み (例えば [4] など) も含まれる．本研究は特に，多くの Wikipedia 記事に配置されている要約的情報である Infobox に着目し，Infobox が担う属性・属性値情報 ([1, 3, 6]) を言語間で対応づけることにより，複数の言語 (日本語，英語，中国語) の Wikipedia Infobox から対訳辞書を抽出する方法について検討する．より具体的には，Infobox を記述するための Infobox テンプレートから属性の言語間対応を抽出し，これらの Infobox テンプレートを引用する Infobox インスタンス群から属性値の言語間対応を抽出する．最終的に，これらの属性・属性値情報の対応付けから対訳辞書を抽出する．

2　Infobox の基本構造

基本的に Wikipedia の記事は，実世界におけるある実体 (entity) を記述するものであり，良く知られているように，Wikipedia の多くの記事は，Infobox と呼ぶ，記述の対象に対する情報を要約した表を配置している．このような Infobox は，対象の記述となる実体のタイプに応じたテンプレート (Infobox テンプレート) を具体化することにより生成されることが多い[1]．

”Infobox_Software” という名称を持つ Infobox テンプレートを引用する日本語のインスタンスの例を図 1 に，これに対応する英語のインスタンスの例を図 2 示す．

Infobox テンプレートは記述対象の実体のタイプに応じて，表として列挙すべき情報項目 (属性) を規定するものであり，具体的な Infobox インスタンスは，特定の実体を具体的に記述するために，属性に対する具体的な情報 (属性値) を供給することによって実現される．また，Infobox テンプレートの多くは，言語間で対応するテンプレートが明示されている[2]．よって対応する Infobox テンプレートにおける属性名の言語間対応付けを行うことが可能となる．さらに，Infobox インスタンスの言語間対応付けを行うことにより，属性値として記述される情報 (その多くはいわゆる固有表現) の言語間対応付けを行うことができる．このように，Infobox の有する情報構造は，多言語の対訳辞書を抽出するための情報源として有望である．

図 1: Infobox インスタンスの例 (日本語)

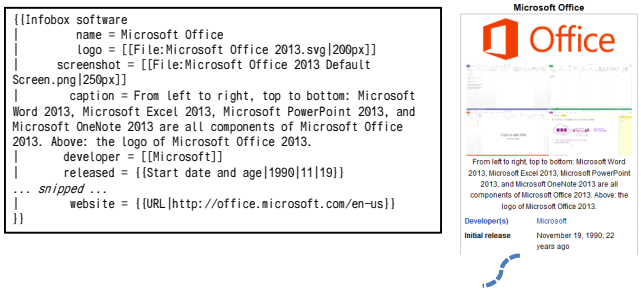

図 2: Infobox インスタンスの例 (英語)

3　Infobox の言語間対応付け

3.1　対応付け処理の概要

図 1，図 2 に例を示したように，各国語版の Infobox テンプレート は，理想的には等価な属性情報をそれぞれの言語で提示する．

[1]2013 年 1 月 3 日現在，例えば日本語では 1,925 種類の Infobox テンプレートが登録されており，最も多く引用されているテンプレート (基礎情報 会社) は，20,422 件の記事から引用されている．

[2]例えば，日本語の Infobox テンプレート 1,925 種類のうち，対応する英語/中国語のテンプレートを持つものは，705/353 件存在する．

これらのソースコードの対比から，開発元:developerなどの属性名の対応付けを行うことができれば，対象とするタイプの実体に関する属性の対応付けが行え，さらに対応付けられた属性のもとで，属性値間の対応付けが可能となるであろうことが推測できる．

しかし実際には，このような属性名の対応は必ずしも自明ではないため，本稿で提案する対応付けは，属性サイドの情報として，後述する属性変数，および，属性名を手がかりとして用いた対応付けを行う一方，属性値を手がかりとする対応付けをも試みる．すなわち，Infobox インスタンスに実際に出現する属性値集合の共通性から逆に属性の対応を求める．最終的に，これらの手がかりに基づく結果を統合することにより，網羅性と精度のバランス良い向上を目指す．

3.2 属性変数を手がかりとする対応付け

記述対象の実体のタイプに応じた各種の Infobox テンプレートも Wiki のソースコードにより記述される．また，これらのテンプレートは，”Infobox” という名前の一つ上位のレベルのテンプレート (以下，テンプレート:Infobox と記述し，各種の Infobox テンプレートと区別する) を引用することが多い．このような場合，各記事で等号の左辺に記述する変数は，テンプレート:Infobox を引用するそれぞれの Infobox テンプレートのソースコードにおいては，テンプレート:Infobox の変数を規定するものとして，等号の右辺の要素として現れる．

英語Wikipediaの“ Template:Infobox airline”のソースコードの一部
```
| label9 = Operating bases
| data9  = {{{bases<includeonly>|</includeonly>}}}
```
日本語Wikiepdiaの“ Template:航空会社情報ボックス”のソースコードの一部
```
| label7 = [[ハブ空港#拠点都市|拠点空港]]
| data7 = {{{拠点空港|{{{bases|}}}}}}
```

図 3: Infobox テンプレートのソースコードの例

航空会社に関する Infobox テンプレートのソースコード例 (一部) を図 3 に示す．ここで例えば，英語版の記述における data9 はテンプレート:Infobox における変数名であり，その右辺の各項目 (例えば”bases”) は，記事において，この Infobox テンプレートを引用する際の変数名として使うことができる．このような項目を本稿では属性変数と呼ぶ．この例において，”bases”という項目が日英間で共通しているように，属性変数は互いに対応する各国語の Infobox テンプレートのソースコードにおいて，共通して現れる (共有属性変数と呼ぶ) 場合があり，確度の高い属性対応付けの手がかりとして利用できる．ただし，項目名 (つまり属性変数名) は文字列として一致するとは限らないため，翻訳して言語を揃えた後に文字列間類似度を求め，これを閾値処理することにより対応関係を認定する．

翻訳処理: 言語間リンクで結ばれている Wikipedia 記事のタイトルから抽出した対訳辞書 (以下，Wiki 対訳辞書; 各言語ペアに対して抽出した対訳数は，英日:432,107，英中:337,411，日中:219,375 件)，および，EDR 電子化辞書[3]，言語グリッド[4]の翻訳サービスの翻訳結果をマージした対訳資源を用いて翻訳を行う．

文字列類似度および閾値処理: Jaro の文字列類似度尺度 [5] を用いる．次節で報告する実験においては，事前実験にもとづき，0.8 を閾値とした．

3.3 属性名を手がかりとする対応付け

図 3 において，label9 (英語) や label7 (日本語) といった変数名の右辺に記載される ”Operating bases” や ”拠点空港” といった項目は，実際に Infobox を表示する際に属性の表示として使われるものであり，本稿ではこれらを属性名と呼ぶ．属性名に関しても，属性変数における場合と同様の手順により対応付けを行う．

3.4 属性値を手がかりとする対応付け

属性変数，属性名を手がかりとする対応付けは，Infobox テンプレート，より厳密にはそのソースコードを対象とする処理であったが，属性値を手がかりとする対応付けにおいては，実際の記事に現れる Infobox インスタンス群を対象とする．

属性値を手がかりとする対応付けの処理概要を図 4 に示す．言語 A，B において互いに対応する Infobox テンプレートを引用する記事群における Infobox から，(ページ，テンプレート，属性変数，属性値) のタプルの集合を抽出し，同一の属性変数ごとに属性値集合を区分する．次に，これらの属性値集合間の等価性を総当り的に評価し，対応する属性の対応付け可否を判定する．本研究では，属性値集合の等価性を判定する際に以下の 2 つの方法を試みたが，いずれの場合も等価性を認めることができる属性値ペアの数をカウントし，これが最大となる属性変数ペアを対応付けとして採用した．

[3] http://www2.nict.go.jp/out-promotion/techtransfer/EDR/J_index.html
[4] http://langrid.org/jp/

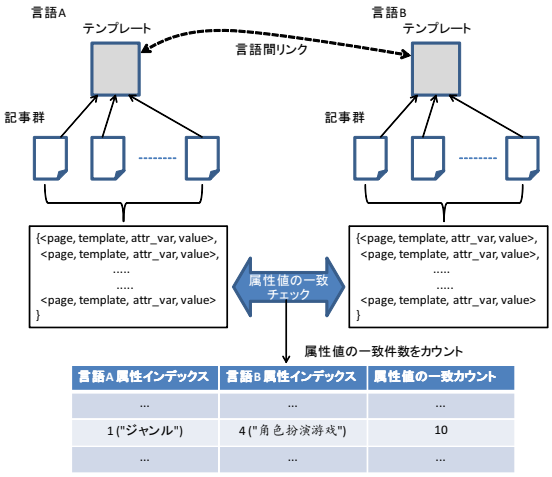

図 4: 属性値を手がかりとする対応付け

- 属性値手法 1: 候補となる属性値ペアの等価性の判定を厳密に行う．より具体的には，翻訳後の文字列が一致する場合に当該の属性値ペアが等価であると判定する．ただし，その属性情報を含む言語 A，言語 B のそれぞれの記事の等価性についてはチェックしない．ここで，記事の言語間等価性は，それらの記事が言語間リンクで結ばれているかどうかに基づく．
- 属性値手法 2: 候補となる属性値ペアの等価性の判定を類似度により行う．より具体的には，前節までと同様の方法により類似度に基づき等価性の判定を行う．本方法では，属性値ペアのデータとしての等価性判定に関してマージンを持たせる一方で，その属性情報を含む言語 A，言語 B のそれぞれの記事の等価性は厳密にチェックする．

3.5 評価実験と結果

評価対象データ: Infobox テンプレートの中でも明示的にテンプレート:Infobox を引用している日本語，英語，中国語の Infobox テンプレートであって，いずれの 2 言語の間においても言語間の対応付けがされている Infobox テンプレートの 3 言語の組 (96 組) を評価対象のデータとし，これらの各 2 言語間の対応付けに関して，正解データを準備した．これらのテンプレートを引用している記事の数はそれぞれ，日本語:102,985 件，英語:788,978 件，中国語:35,875 件であった．

結果の統合: 結果の統合には様々な方法が考えられるが，本報告の実験では，属性変数による結果をベースとし，これを属性名による結果で補完する手法 (統合手法 1)，この結果をさらに属性値による結果で補完する手法 (統合手法 2) を比較する．なお，統合手法 2 では，まず属性値手法 1 による結果を補完し，さらに属性値手法 2 による結果を補完する．

表 2: 対訳エントリの抽出件数と分類

言語ペア	np	n1	n2	n3 (n3/n2 比率)
英日	210	31,053	9,735	8,265 (84.8%)
英中	165	36,970	10,869	9,390 (86.4%)
日中	152	37,627	16,749	14,433 (86.2%)

評価結果と考察: 表 1 に，適合率 (P)・再現率 (R)・F 値による評価結果を示す．まず，属性変数を手がかりとする属性の対応付けは良好な結果を示している．これは，今回の対象の Infobox テンプレートにおいては，共有属性変数を持つ属性ペアが多く存在した (93.8%) ことが原因である．次に，統合手法 1，統合手法 2 の再現率が属性変数のみを用いる場合よりも多少高いことから，これらの手がかりは，属性変数が存在しないケースを補完するために有効であるが，さらなる精度向上が必要であるとことが分かる．また，統合手法 1 の方が統合手法 2 よりも総合的に優れているため，属性値の対応付けの精度や結果の統合の仕方が現状では十分ではないと考えられる．属性の対応付けの決定を補完する情報として用いるならば，属性値の等価性まで判定する必要はなく，データタイプの一致でも有用である可能性がある．

4 対訳辞書の抽出実験と評価

以上の結果に基づき，統合手法 1 を用いて属性の対応付けを行った後，その制約のもとで属性値を対応付けるという手順により，対訳辞書の抽出実験を行った．以下では抽出された対訳エントリの量と質を評価する．

この実験で用いた各言語ペアにおける Infobox テンプレートペアの数 (np)，抽出された対訳エントリの件数を表 2 に示す．ここで，n1 とは抽出されたエントリの総数，n2 はノイズや重複を排除した後のエントリ数，n3 はその中で Wiki 対訳辞書に存在しないエントリ数を示す．このように，抽出エントリ件数の絶対量はさほど多くないものの，抽出されたエントリには比較的高い割合 (表中の n3/n2 比率) で新規のエントリが得られていることが分かる．

次に，各言語ペアに対して抽出した新規エントリから，生起頻度が上位のもの 100 件 (A 群[5])，および，

[5]各言語ペアにおける頻度の最大値/平均値/最頻値は，英日:512/36.9/11，英中:121/18.8/8，日中:334/32.2/13 であった．

表 1: 属性対応付けの評価結果

	属性変数のみ利用			統合手法 1			統合手法 2		
言語ペア	P	R	F	P	R	F	P	R	F
英日	0.953	0.918	0.936	0.952	0.916	0.935	0.857	0.937	0.896
英中	0.963	0.954	0.959	0.963	0.964	0.960	0.924	0.960	0.942
日中	0.940	0.941	0.940	0.939	0.950	0.944	0.874	0.960	0.915

表 3: 対訳エントリの品質評価 (A 群/B 群)

言語ペア	○	△ 1	△ 2	×
英日	44/23	1/4	41/11	14/62
英中	45/38	2/3	34/5	19/54
日中	79/48	1/4	10/7	10/41

生起頻度が 1 のものからランダムに選んだ 100 件 (B 群) について，対訳としての品質を評価した．この結果を表 3 に示す．ここで，○とはそのまま対訳として利用できるレベル，△ 1 とは文字列の一部にノイズを含むもの，△ 2 とは日本語・中国語側が英語表記，あるいは，対訳が数字列などの有用性の低いもの，×は対訳誤りである．予想されるように B 群の抽出精度は A 群より劣る．また誤りには，不正な Wikipedia ソースコードが原因のものも含まれる．

前節の抽出数の評価とこの結果を合わせると，テンプレートペアあたりで抽出される正しい対訳エントリの平均数は，抽出精度を B 群の実績値により低く見積もった場合において，英日で 9.1 件，英中で 21.6 件，日中で 45.6 件となる．現時点でこの件数の多少については議論できないが，少なくとも日中に対する対訳抽出が良好である傾向が確認できる．これは両者の対訳が類似度の高い漢字文字列による場合が多いことを反映している．

また，抽出に成功した対訳のタイプについて調べてみると，日中に関しては，人名・地名の占める割合が比較的高い (A 群で 68.4%，B 群で 77.8%) という特徴がみられた．なかでも，欧米人の人名の表記は日本語ではカタカナで行われることが多く，これに対する漢字表記による中国語の対訳を良好に抽出できている．例えば,「リチャード・マークス」に対しては,「李察・馬克斯」および,「理察・馬克斯」が抽出されており，漢字表記のヴァリアントの抽出に有用である．以上より提案手法は，記事中の Infobox では言及されるものの，独立した記事が (まだ) 作成されていない実体に対する対訳の抽出に有効である．また，実体を指す固有表現以外の語に関しても，分野・ジャンルに固有の訳語を抽出できる場合がある (例: rider に対して選手 (ジャンル=自転車競技))．

5 おわりに

Wikipedia Infobox が担う属性・属性値情報を言語間で対応づけることにより，日本語，英語，中国語を対象とする対訳辞書を抽出する方法について報告した．特に，従来研究においてほとんど注目されることのなかった共有属性変数が対応付けにおいて有用であることを示したが，共有属性変数は全ての Infobox テンプレートにおいて存在するわけではなく，また，言語・対象の実体のタイプに対する偏りも大きい．このため，やはり属性名・属性値に基づく対応付けの精度と適用範囲を高めていく必要がある．

抽出される対訳辞書の量・質は，属性の対応付けが改善されることにより，向上することが期待できる．一方で，DBpedia のようなクリーニングされたデータが各国語で整備されれば，それを利用することにより不正な Wikipedia ソースコードによる失敗を避けることが可能と考えられる．

参考文献

[1] Adar, E., et al. Information arbitrage across multilingual Wikipedia. *Proc. of WSDM '09*, pp.94–103.

[2] Bizer, C., et al. (2009). DBpedia - A crystalization point for the Web of data. *Journal of Web Semantics*, Volume 7, Issue 3, pp.154–165.

[3] Bouma, G., et al. (2009). Cross-lingual alignment and completion of Wikipedia templates. In *Proc. of CLIAWS3*, pp.21–29.

[4] Erdmann, M., et al. (2008). Improving the extraction of bilingual terminology from Wikipedia. *ACM Trans. on TOMCAAP*, Volume 5, Issue 4, No. 31.

[5] Jaro, M. (1995). Probabilistic linkage of large public health data file. *Statistics in Medicine* 14 (5-7): 491–498.

[6] Nguyen, T., et al. (2012). Multilingual schema matching for Wikipedia Infoboxes. *Proc. of VLDB 2012*, pp.133–144.

[7] Wang, Z., et al. Cross-lingual knowledge linking across Wiki knowledge bases. *Proc. of WWW 2012*, pp.459–468.

[8] 柴木優美, 永田昌明, 山本和英. カテゴリ名と記事名の意味属性分類に基づく Wikipedia からの上位下位オントロジーの構築. 自然言語処理, Vol.19, No.4, pp.229–279.