中間言語を用いたインドネシア語-日本語対訳辞書の拡充

Expanding Indonesia-Japanese Translate Dictionary Using Pivot Language

脇田 敏行 *,　　土屋 雅稔 †,　　中川 聖一 *

豊橋技術科学大学

* 情報工学系 / † 情報メディア基盤センター

1. はじめに

言語横断情報検索や言語横断質問応答，機械翻訳などの 2 つの言語に関わる処理を実現するには，その言語対に対する大規模対訳辞書などの言語横断言語資源が必要である．情報流通技術の発達に伴って，様々な言語で記述された情報を活用することが可能となりつつあり，複数言語を対象とする自然言語処理技術はますます重要な課題となることが予想される．しかし，世界には数多くの言語が存在するため，あらゆる言語対を対象として豊富な言語資源を整備することは，非現実的である．現実には，需要の大きい一部の言語対については大規模な言語資源が利用できるが，それ以外の多くの言語対については，小規模な対訳辞書しか利用できない場合が多い．

このような状況に対応するために，対訳辞書の自動構築や，訳語獲得などの研究が広く行われてきた．田中ら [1] は，英語を中間言語として利用して，和仏対訳辞書を作成する方法を提案している．この方法では，和英辞書と英仏辞書を利用して，日本語単語に対するフランス語訳語候補を獲得し，仏英辞書と英和辞書を利用して，得られたフランス語単語に対する日本語訳語候補を調べる（逆引きを行う）ことによって，訳語候補の絞り込みを行い，訳語推定精度を改善している．この方法で，名詞を対象とした場合の精度は 76%，再現率は 44%である．張ら [2] は，英語を中間言語として利用して，日中対訳辞書を作成する方法を提案している．この方法では，和英辞書と英中辞書を利用して，日本語単語に対する中国語訳語候補を獲得し，日本語と中国語の品詞情報と漢字情報を利用して訳語候補の順位付けを行っている．この方法で，第 1 位に順位付けられた訳語候補のみを出力した場合の精度は 81.4%である．これらの研究では，いずれも，対象としている言語対そのものに対する対訳辞書の情報は利用していない．それに対して，何らかの対訳辞書が存在することを前提として，その対訳辞書に収録されていない未知語の訳語を推定する研究も非常に多い．例えば，宇津呂ら [3] は，既存の大規模対訳辞書と大規模対訳コーパスを利用して，新語に対する訳語を獲得する方法を提案している．この方法では，ウェブ上の大規模対訳コーパスから新語に対する訳語候補を獲得し，訳語候補に対する単語共起ベクトルと未知語に対する単語共起ベクトルとの類似度を求めて，この類似度の高いものを訳語として出力している．この研究は，新語の訳語を獲得するという目的で行われているため，対象とする言語対以外の言語対に対する対訳辞書の情報を利用していない．

本稿では，規模を問わなければ，かなり多くの言語対について対訳辞書が利用できる状況を考慮して，既存の小規模な対訳辞書を，その言語対の各言語と中間言語に関する言語資源を利用して，大規模な対訳辞書に拡充する方法を提案する．実際に，提案手法を用いて，インドネシア語-日本語辞書を拡充した結果について報告する．

2. 中間言語を用いた対訳辞書の拡充

2.1 対訳辞書の拡充

ある入力言語から，ある出力言語への対訳辞書を作成するとき，以下のような状況を仮定する．

> **仮定:** ある中間言語を考えると，入力言語から出力言語への小規模な対訳辞書 (以下，**種辞書**と呼ぶ) と，入力言語から中間言語への大規模な対訳辞書および中間言語から出力言語への大規模な対訳辞書が存在する．

このような仮定は，英語を中間言語とすると，かなり多くの言語対に対して成り立つことが期待できる．

本研究では，上述の仮定の下で，入力言語から中間言語への大規模対訳辞書には登録されているが，種辞書には登録されていない語の訳語を推定するというタスク（対訳辞書の拡充）を扱う．

中間言語を利用して対訳辞書を作成する先行研究とは異なり，本研究では，種辞書の情報を有効に活用して，より大規模な対訳辞書を作成することに焦点をあてている．また，未知語（新語）に対して訳語を獲得する場合，未知語（新語）の多くは名詞であるから，名詞のみを対象とする手法で十分である．それに対して，小規模な種辞書を拡充する場合には，種辞書には収録されていない動詞・形容詞に対する対応が必要になる．この詳細は 3.2 節で述べる．

2.2 中間言語と共起情報を用いた拡充方法

本研究で提案する拡充方法は，以下の 2 段階からなる．

(1) 入力言語のコーパスを用いて，翻訳したい単語と，種辞書に登録されている見出し語の単語共起ベクトルを作成する．次に，その単語共起ベクトルを，種辞書を用いて，出力言語上のベクトルに変換する．

(2) 入力言語から中間言語への対訳辞書と，中間言語から出力言語への対訳辞書を利用して，訳語候補を列挙する．出力言語のコーパスを用いて，訳語候補それぞれの単語共起ベクトルを作成し，前段階で得られたベクトルとの類似度を用いて，訳語

を決定する．

最初に，入力言語上の単語共起ベクトルを，種辞書を用いて，出力言語上のベクトルに変換する．言語を問わず，コーパス上における単語 w_i, w_j の共起頻度は $f(w_i, w_j)$ と表す．種辞書 D の全見出し語を $x_i (i = 1, 2, \ldots, n)$ とすると，入力言語の単語 x_s の共起ベクトル $\mathbf{v}(x_s)$ は，次式のように表される．

$$\mathbf{v}(x_s) = (f(x_s, x_1), \ldots, f(x_s, x_n)) \quad (1)$$

つまり，この共起ベクトル $\mathbf{v}(x_s)$ の各次元は，入力言語の単語と対応している．この共起ベクトル $\mathbf{v}(x_s)$ を，種辞書を用いて，各次元が出力言語の単語と対応するようなベクトル $\mathbf{v}_t(x_s)$ に変換する．

$$\mathbf{v}_t(x_s) = (f_t(x_s, z_1), \ldots, f_t(x_s, z_m)) \quad (2)$$

ここで，$z_j (j = 1, 2, \ldots, m)$ は，種辞書に現れる全ての訳語である．また，$f_t(x_s, z_k)$ は，単語 x_s に関する入力言語コーパス上の共起頻度を，出力言語の単語 z_j との共起の程度を示すように変換する関数であり，次のように定義する．

$$f_t(x_s, z_j) = \sum_{i=1}^{n} f(x_s, x_i) \cdot \delta(x_i, z_j) \quad (3)$$

ここで，$\delta(x_i, z_j)$ は，単語 z_j が単語 x_i の訳語であるかどうかを示す関数であり，単語 x_i を種辞書で調べたときに得られる訳語集合を $D(x_i)$ とすると，次式によって表される．

$$\delta(x_i, z_j) = \begin{cases} 1 & \text{if } z_j \in D(x_i) \\ 0 & \text{otherwise} \end{cases} \quad (4)$$

次に，訳語候補を，以下の手順で列挙する．入力言語の単語 x_s について，入力言語から中間言語への対訳辞書を検索して，中間言語の説明文集合 $\mathbf{Y}_s$ を得る．得られた説明文 $\mathbf{y}_s \in \mathbf{Y}_s$ を用いて，中間言語から出力言語への対訳辞書を検索し，出力言語の説明文候補の集合 $\mathbf{Z}_s$ を得る．説明文 $\mathbf{y}_s$ が単語 1 語からなっていた場合は，そのまま中間言語から出力言語への対訳辞書を検索し，説明文 $\mathbf{y}_s$ が複数の語からなっていた場合は，構成語全てを用いて，中間言語から出力言語への対訳辞書を検索する．続いて，説明文候補集合 $\mathbf{Z}_s$ に含まれる説明文候補 $\mathbf{z}_s = z_s^1 z_s^2 \cdots z_s^l$ について，次式のように表される共起ベクトルを求める．

$$\mathbf{u}(\mathbf{z}_s) = \left(\sum_{k=1}^{l} f(z_s^k, z_1), \ldots, \sum_{k=1}^{l} f(z_s^k, z_m) \right) \quad (5)$$

ベクトル $\mathbf{v}_t(x_s)$ とベクトル $\mathbf{u}(\mathbf{z}_s)$ の cosine 類似度 $s(\mathbf{v}_t(x_s), \mathbf{u}(\mathbf{z}_s))$ を計算し，適当な条件を満たした説明文候補 $\mathbf{z}_s$ を，単語 x_s の説明文（訳語）として出力する．本研究では，条件として，(1) 類似度の大きい説明文から順に出力する，(2) 類似度が適当な閾値より大きい説明文を出力する，という 2 通りの方法を考える．この評価については，3.3 節で述べる．

表 1 コーパスの諸元

諸量	インドネシア語	日本語
記事数	71099	304329
文数	1333919	3454204
のべ語数	10369324	139142566
異なり語数	17169	118080

式 (1) および式 (5) を用いて共起ベクトルを求めるとき，単純な共起頻度 $f(w_i, w_j)$ を用いる代わりに，適当な補正を加える方法が考えられる．本研究では，以下の 2 通りの補正した共起頻度を用いる方法を検討する．

$$f_{\mathrm{IDF}}(w_i, w_j) = \frac{f(w_i, w_j)}{df(w_j)} \quad (6)$$

$$f_{\mathrm{TFIDF}}(w_i, w_j) = \frac{f(w_i, w_j) \cdot tf(w_j)}{df(w_j)} \quad (7)$$

ここで，$tf(w)$ は，ある単語 w の単語出現頻度であり，$df(w)$ は，ある単語 w の文書出現頻度である．これらの補正方法の評価については，3.4 節で述べる．

3. 実　　験

入力言語をインドネシア語，中間言語を英語，出力言語を日本語として，対訳辞書の拡充を行った実験について報告する．

3.1 実験条件

本研究では，日本語コーパスとして，毎日新聞 CD–ROM(1993 年 ~1995 年) を，MeCab☆で形態素解析したデータを用いた．共起ベクトルを用いて単語間の意味的な類似度を測定するには，なるべく類似したドメインに対するコーパスの方が良い結果が得られると予想される．しかし，インドネシア語に対する既存の言語資源は大変少ないため，インドネシア国内向けに編集・公開されているウェブ新聞☆☆ の記事を，インドネシア語コーパスとして用いた．各コーパスの諸元は**表 1** の通りである．

対訳辞書の拡充手法を評価するには，インドネシア語から英語への対訳辞書 [4] に収録されているが，種辞書には収録されていない語，つまり実際の拡充対象となる語を対象として，各種評価を行うべきである．しかし，そのような語については正解訳語のリストが存在せず，安定した評価が難しい．そのため，本研究では，インドネシア語から日本語への対訳辞書 [5] から，500 個の訳語対をテスト用として取り出し，残りの訳語対のみを登録した辞書を種辞書として実験を行う．この時，インドネシア語コーパスの一部である開発用コーパス☆☆☆を用いて，拡充対象となる語の頻度分布を調査し，得られた頻度分布と，テスト用訳語対のインドネシア語単語の頻度分布が概ね等しくなるように，テスト用訳語対を選択した．結

☆ http://chasen.org/~taku/software/mecab/

☆☆ http://www.kompas.com/
http://www.tempointeraktif.com/

☆☆☆ インドネシア語コーパスは，ウェブで公開されている新聞記事を収集して作成したため，研究開始当初は記事数がまだ少なかった．

表 2 インドネシア語-日本語の訳語対のインドネシア語コーパス中の頻度分布

頻度 f	インドネシア語-英語辞書		テスト用訳語対		インドネシア語-日本語辞書	
$1000 \leq f$	1037	(11.6%)	58	(10.2%)	154	(3.6%)
$500 \leq f < 1000$	1357	(15.2%)	76	(15.2%)	305	(7.1%)
$200 \leq f < 500$	1057	(11.9%)	61	(12.1%)	340	(7.9%)
$100 \leq f < 200$	1055	(11.8%)	60	(12.0%)	426	(9.9%)
$50 \leq f < 100$	1280	(14.4%)	72	(14.4%)	654	(15.2%)
$20 \leq f < 50$	823	(9.2%)	48	(9.6%)	513	(11.9%)
$10 \leq f < 20$	707	(7.9%)	41	(8.2%)	494	(11.5%)
$5 \leq f < 10$	731	(8.2%)	42	(8.4%)	607	(14.1%)
$2 \leq f < 5$	395	(4.4%)	22	(4.4%)	352	(8.2%)
$f = 1$	467	(5.2%)	27	(5.4%)	452	(10.5%)

表 3 語彙数と品詞分布

品詞	インドネシア語-日本語辞書		インドネシア語-英語辞書	
名詞	4085	(57.4%)	15718	(53.5%)
動詞	1910	(26.8%)	9600	(32.7%)
形容詞	795	(11.2%)	3390	(11.5%)
その他	330	(4.6%)	682	(2.3%)
計	7120	(100%)	29390	(100%)

果ぁを**表 2** に示す．インドネシア語 1 語に対応する正解の日本語訳語は平均 1.38 個である．また，英和辞書としては英辞郎 [6] を用いた．

評価尺度としては，次式によって定義される精度，再現率，訳語含有率を用いた．

$$\text{精度} = a/b$$
$$\text{再現率} = c/d$$
$$\text{訳語含有率} = e/f$$

ただし，出力された候補の内，正解と判定された候補の数を a，出力された候補の総数を b とする．また，正解の内，出力された正解の数を c，正解の総数を d とする☆．e は正解訳語が 1 つ以上見つかったテスト単語の数，f はテスト単語の総数 (500) である．

3.2 対象となる訳語対の品詞別分類

インドネシア語から日本語への対訳辞書に収録されている見出し語の品詞別分類と，インドネシア語から英語への対訳辞書に収録されている見出し語の品詞別分類を**表 3** に示す．表より，2 つの辞書の品詞別分類に大きな差はなく，約 7,000 語から約 30,000 語にインドネシア語の語彙が拡大するとき，名詞ばかりが増えるのではなく，動詞・形容詞についてもほぼ均等に増加していることが分かる．そのため，小規模な種辞書を拡充する場合には，未知語（新語）に対する訳語獲得とは異なり，名詞だけではなく，動詞や形容詞についても訳語を推定する必要がある．

インドネシア語から日本語への対訳辞書に収録されている訳語対を，インドネシア語単語の品詞と日本語訳語の品詞によって分類した結果を**表 4** に示す．インドネシア語から日本語への訳語対を獲得する場合，名詞を対象とする場合には，名詞から名詞への訳語対応のみを考えれば良いが，動詞・形容詞を対象とする場合は，説明文に対応づけられる場合を考慮する必要があることが分かる．

☆ 正解と同義の表現が出力された場合は，人手で判定を行った．そのとき，1 つの正解に対して複数の出力が対応付けられ，a と c が等しくならないことがある．

表 6 共起頻度の補正方法による比較

尺度	補正なし	式 (6)	式 (7)
精度	46.3%	42.7%	46.5%
再現率	60.8%	56.4%	61.5%
訳語含有率	74.2%	68.8%	74.4%

上位 3 候補を出力した場合の精度，再現率，訳語正解率を示した．

3.3 出力する訳語の選択方法による比較

類似度によって整列された訳語候補リストから，どの部分を訳語候補として出力するかを検討する．上位 n 語を取り出した場合と，類似度 s が閾値より大きい語を取り出した場合の精度・再現率・訳語含有率を**表 5** に示す．

表 5 より，類似度と閾値を比較して出力する訳語を選択すると，低い閾値を用いた場合には，良い精度が得られず，高い閾値を用いた場合には，精度は改善されるが，出力される訳語が極端に少なくなってしまう問題がある．したがって，この提案手法に対しては，適当な閾値と比較して出力する訳語を選択する方法は適していないことが分かる．

3.4 共起頻度の補正方法による比較

単純な共起頻度を用いて共起ベクトルを求めた場合，文書出現頻度を用いて式 (6) のように補正して共起ベクトルを求めた場合，単語出現頻度と文書出現頻度を用いて式 (7) のように補正して共起ベクトルを求めた場合を比較した．結果を**表 6** に示す．表より，これらの補正による変化は殆んど観察されず，単純に共起頻度を用いて共起ベクトルを求める方法が良いようである．

3.5 品詞別の比較

テスト用の訳語対集合を，インドネシア語の品詞によって名詞・動詞・形容詞の 3 つに分類し，それぞれに対して訳語推定を行った場合の結果を**表 7** に示す．名詞・形容詞に対する精度・再現率・訳語含有率は，ほとんど変わらないが，動詞に対する精度・再現率・訳語含有率はわずかに低くなっている．これは，動詞に対する正解訳語候補が，名詞・形容詞に対する正解訳語候補よりも多いことが原因と考えられる．

3.6 頻度別の比較

テスト用の訳語対集合を，インドネシア語の開発用コー

表 4 インドネシア語-日本語辞書における品詞変化

		日本語			
		名詞	動詞	形容詞	その他
インドネシア語	名詞	**4381 (77.2%)**	11 (0.2%)	10 (0.2%)	1273 (22.4%)
	動詞	58 (2.2%)	1300 (48.7%)	7 (0.3%)	**1306 (48.9%)**
	形容詞	249 (20.7%)	86 (7.2%)	131 (10.9%)	**736 (61.3%)**
	その他	2656 (27.6%)	848 (8.8%)	11 (0.1%)	**6097 (63.4%)**

「その他」には，品詞が不明だった場合と，1 語以上からなっていた場合が含まれている．

表 5 閾値の指定方法による比較

	上位 n 候補を出力					類似度 $s >$ 閾値 x である候補を出力				全候補を出力
尺度	$n=1$	$n=2$	$n=3$	$n=5$	$n=10$	$x=0.05$	$x=0.1$	$x=0.2$	$x=0.3$	(ベースライン)
精度	54.6%	50.9%	46.3%	40.3%	33.0%	20.5%	23.0%	30.8%	43.7%	20.2%
再現率	41.4%	53.9%	60.8%	67.3%	75.3%	75.0%	60.3%	31.7%	7.3%	23.8%
訳語含有率	54.6%	68.8%	74.2%	79.8%	85.8%	82.2%	65.4%	34.2%	8.8%	23.8%

表 7 品詞による比較

尺度	名詞	動詞	形容詞
精度	49.5%	40.1%	47.3%
再現率	66.1%	51.2%	60.3%
訳語含有率	77.7%	66.7%	76.0%

上位 3 候補を出力した場合の精度，再現率，訳語正解率を示した．

表 8 頻度による比較

頻度 f	精度	再現率	訳語含有率
$1000 \leq f$	37.7%	44.0%	74.1%
$500 \leq f < 1000$	43.9%	57.9%	77.2%
$200 \leq f < 500$	54.7%	58.6%	73.8%
$100 \leq f < 200$	46.9%	53.3%	65.9%
$50 \leq f < 100$	57.8%	70.8%	87.5%
$20 \leq f < 50$	39.6%	57.3%	66.7%
$10 \leq f < 20$	45.7%	68.1%	75.0%
$5 \leq f < 10$	45.3%	65.3%	77.0%
$2 \leq f < 5$	42.2%	62.2%	71.1%
$f = 1$	50.8%	67.8%	78.4%

パス上の頻度によって分類した場合の結果を，**表 8** に示す．頻度が精度・再現率・訳語正解率に与えている影響は，それほど大きくないようである．したがって，本来の目的である種辞書中に存在しない単語に対する訳語も，同程度の精度で得られると期待できる．

3.7 種辞書の大きさによる比較

次に，種辞書の大きさが，訳語推定に対して与える影響について検討する．種辞書の訳語対を，インドネシア語の開発用コーパス上の頻度順によって整列し，上位 n 対のみを残すことによって，n 対からなる小規模な種辞書を作成した．この種辞書を用いて，テスト用訳語対に対する訳語推定を行った場合の精度変化を**図 1** に示す．図より，訳語推定を行うには 3000 語程度の種辞書が必要と考えられる．

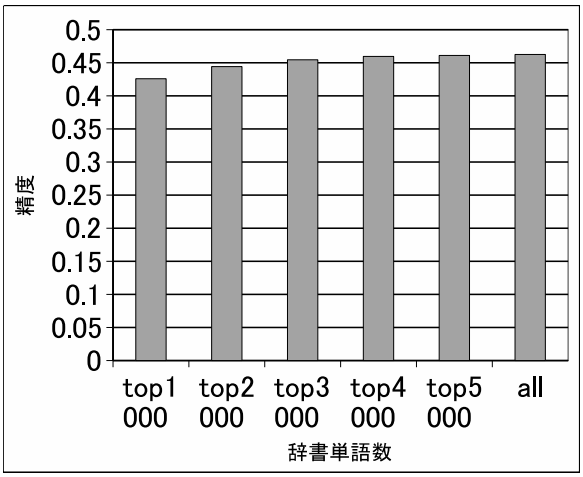

図 1 種辞書の大きさによる精度の変化

4. むすび

本稿では，インドネシア語-英語辞書および英語-日本語辞書を利用して訳語候補を取り出し，インドネシア語コーパスと日本語コーパスの共起情報を用いて訳語候補の絞り込みを行って，小規模なインドネシア語-日本語辞書を拡充する方法を提案した．提案手法を用いて実際に辞書を拡充したところ，精度 54.6%で拡充することができた．今後は，インドネシア語から日本語への言語横断情報検索システム [7] に組み込むことにより，本提案手法の有効性を確認したいと考えている．

参考文献

1) 田中久美子, 梅村恭司, 岩崎英哉: 第 3 言語を介した対訳辞書の作成, 情報処理学会論文誌, Vol. 39, No. 6, pp. 1915–1924 (1996).
2) 張玉潔, 馬青, 井佐原均: 英語を介した日中対訳辞書の自動構築, 自然言語処理, Vol. 12, No. 2, pp. 63–85 (2005).
3) 宇津呂武仁, 日野浩平, 堀内貴司, 中川聖一: 日英関連報道記事を用いた訳語対応推定, 自然言語処理, Vol. 12, No. 5, pp. 43–69 (2005).
4) Agency for The Assessment and Application of Technology: Kamus Elektornik Bahasa Indonesia. `http://nlp.aia.bppt.go.id/kebi`.
5) Sanggar Bahasa Indonesia Proyek: KMSMINI2000. `http://m1.ryu.titech.ac.jp/~indonesia/todai/dokumen/kamusjpina.pdf`.
6) 道端秀樹 (編): 英辞朗, アルク (2002).
7) Purwarianti, A., Tsuchiya, M., Nakagawa, S.: Query Transitive Translation Using IR Score for Indonesian-Japanese CLIR, *Proceedings of Second Asia Information Retrieval Symposium (AIRS2005)*, pp. 565–570 (2005).