異言語の語彙概念の対応付けのための手がかり情報の有効性評価

林 良彦
大阪大学 大学院言語文化研究科
hayashi@lang.osaka-u.ac.jp

1　はじめに

独立に構築された異なる言語の異なる語彙資源における語彙概念を相互に対応付ける「異言語の語彙概念の対応付け」の検討を進めている．ここで対象とする対応付けは，しばしば alignment と称される語彙概念体系全体を整合的に対応付けようとするものではなく，語彙概念ノードの間の局所的な対応付け (alignment に対して matching と呼ばれる [3]) を行うものである．このような局所的な対応付けを比較的軽量な処理により実現できれば，Web サービスによる動的な実行環境への適用が可能となり，異種の言語資源を Web サービス上で組み合わせることによる仮想的な言語資源の動的な実現に寄与しうる．

以上のような動機のもと，原言語 (SL) と目的言語 (TL) の語彙概念を規定する同義語集合どうしの対応の度合いを評価することによる異言語の語彙概念の対応付け手法を提案し，その有効性を調べた [4], [5]．その結果，目的言語における語義タグ付きコーパスが有効な手がかりとなることを確認した．本稿ではさらに，語彙資源において与えられている語彙概念に対する定義・説明 (gloss) テキストの類似性が対応付けの手がかりとしていかに有効に利用できるかを検討する．ここで注意すべきことは，類似性を評価すべき gloss テキストの言語が異なっていることである．今回は，Web サービスとして提供されている翻訳機能を適用し，同一言語上でテキスト類似性の評価を行った．

2　タスク設定と手がかり情報

本研究におけるタスクは，原言語 SL(例えば日本語) の語彙概念体系における語彙概念 s に対して，意味的に対応しうる目的言語 TL(例えば英語) における語彙概念 t を求めることである．

ここで語彙概念体系としては，Princeton WordNet [7] (以下，PWN) に準拠する情報構造を持つ語彙資源を対象とする．すなわち，語彙概念は同義語集合により規定されるノードであり，これらの間の語彙概念関係により意味ネットワークが形成される．本研究の評価実験では，このような要請を満たす日本語の語彙概念体系として，日本語 WordNet (以下，WN-Ja)[2]，および，EDR 電子化辞書 (以下，EDR)[8] を用いる．WN-Ja は基本的には PWN を構造を保持したまま日本語に翻訳したものである．一方，EDR の情報構造は PWN と同様の語彙概念体系とは言えないが，EDR を構成する各辞書のエントリは概念識別子を有していることから，同一の概念識別子を持つ語の集合は PWN 同様の同義語集合 (擬似 synset) とみなすことができる．

上記のタスクへのアプローチとしては，対訳資源を用いて対応先の候補となる TL における語彙概念集合 $\{t_i\}$ を求め，各候補 t_i に対して意味的な関連度 $score(s, t_i)$ を計算する方法をとる．ここで $score(s, t)$ を計算するための手がかりとしては様々な情報が考えられるが，[4], [5] では，s，t を規定する同義語集合の言語横断的な対応度のみを用いる手法の評価を行い，特に対訳資源により仲介される SL における単語とターゲットとする語彙概念との言語横断的な関連度を計量するという目的において，目的言語における語義タグ付きコーパス (より具体的には Princeton Annotated Gloss Corpus[1]) が有用であることを示した．本報告では，同義語集合の情報に加え，gloss テキストが手がかり情報としていかに有効に利用できるかを検討する．

3　意味的関連度 $score(s, t)$ の計算

SL における語彙概念 s と TL における語彙概念 t の間の意味的関連度 $score(s, t)$ を以下のように定式化する．ここで，$score_p(s, t)$ は同義語集合の対応度に基づく関連度，$score_g(s, t)$ は gloss テキストの類似性に基づく関連度，$0 \leq \beta \leq 1$ は実験的に定める重みである．

$$score(s, t) \equiv (1 - \beta) scope_p(s, t) + \beta score_g(s, t) \quad (1)$$

[1] http://wordnet.princeton.edu/glosstag.shtml

3.1 $score_p(s,t)$ の計算

$score_p(s,t)$ を以下のように定式化する．

$$score_p(s,t) \equiv \sum_{x_i \in \sigma(s)} \omega(x_i,s,t) \times score'(x_i,t) \quad (2)$$

ここで，$\sigma(s)$ は s の同義語集合を表し，$\omega(x_i,s,t)$ は，次式で定義する $score'(x_i,t)$ に対する重み関数である．[4] では，実験的にこの関数形を定めているが，本稿ではその詳細は省略する．

$$score'(x_i,t) \equiv p(t|x) = p(t) \sum_{y_j \in \tau(x)} \frac{p(y_j|t)p(y_j|x)}{p(y_j)} \quad (3)$$

式 (3) は，SL 同義語集合の各要素 x_i と対応付けの候補となっている TL 語彙概念 t との関連度をコーパス統計量から計算する．ここで，$\tau(x)$ は SL の語 x に対する訳語集合である．また，$p(y_i|t)$ は語彙概念 t が与えられたときの TL の語 y_i の事後確率であり，TL における語義タグ付きコーパスから最尤推定する．さらに，$p(y_j|x)$ は翻訳確率であり，複数の対訳辞書におけるエントリ，および，パラレルコーパス[2]から推定する．式 (3) の導出などは，[4] を参照されたい．

3.2 $score_g(s,t)$ の計算

$score_g(s,t)$ を計算するためには，gloss テキストの言語横断的な類似性を評価する必要がある．今回は最も簡便な方法として，SL 側の gloss テキストを翻訳して言語を TL に揃えた状況でテキスト類似性を評価することとし，$score_g(s,t)$ を以下のように定式化する．

$$score_g(s,t) \equiv TextSim(\tau_s(gloss(s)), gloss(t)) \quad (4)$$

ここで，$gloss(s)$ は語彙概念 s の gloss テキスト，$\tau_s(a)$ は SL におけるテキスト a を TL に翻訳した結果を表す．今回，$\tau_s(a)$ は言語グリッドにおいて Web サービス[3]として提供されている日英翻訳機能から 2 つのもの (MT-A，MT-B と称する) を利用した．そこで，それぞれによる翻訳結果を $\tau_A(a)$，$\tau_B(a)$ と表す．また，両者の結果を連結したものを $\tau_C(a)$ と表す．

一方，$TextSim(u,v)$ は同一言語のテキスト u，v の間の類似性を定量化する関数である．単一言語における語義曖昧性解消を目的とした従来研究 [6, 1] では，単語系列のオーバーラップを利用しているが，本研究では一方のテキストが翻訳されたものであるため，単語系列の情報が有効に働く期待は低い．そこで，lemmatize やストップ語の除去などの正規化を行った後のそれぞれを bag-of-words に基づく単語集合 (U, V) とみなし，集合間の類似度測度を適用する．

[2]synset ごとに対応付けられた PWN と WN-Ja の gloss テキストから構成できる．

[3]http://langrid.org/service_manager/language-services

4 評価実験

4.1 概要

評価実験のタスク: 以下の 2 つを設定する．以下では処理対象となる SL の語彙概念をクエリと呼ぶ．

- WN-Ja からの復元 (recovery) タスク: 与えられた WN-Ja のクエリから，あらかじめ対応していることが分かっている PWN の synset を見つける．
- EDR からの発見 (discovery) タスク: 異なる言語の異なる語彙概念体系の間の対応付けという本研究の目的に即したタスクであり，与えられたクエリ (EDR の擬似 synset) に意味的に関連しうる PWN の synsets を求める．

WN-Ja からの復元タスクにおいては，PWN と WN-Ja との体系全体にわたる対応関係が存在するため，正解データを人手で作成する必要がない．今回の実験では，Core WordNet[4]に含まれる PWN synset に対応する WN-Ja の synset をクエリ集合とした．

一方，EDR からの発見タスクに関しては，[4] に詳細を示す手順により，203 個の EDR 擬似 synset (日本語単語のみ利用) からなるクエリ集合に対して正解データを作成した．ここで，対応付け候補の PWN synset に対しては，同義レベル (Syn-level) の対応，何らかの関係があるレベル (Rel-level) の 2 つの適合レベルを設けた．なお，EDR では PWN と異なり，語彙概念体系を品詞により区分しない．よって適合性の評定においては，品詞の異なりを許容している．また，これらの適合性に合致する PWN synset はクエリあたり複数ありえるが，以下の評価においては，これらの中の最上位に得られたものだけを計数の対象とした．

評価尺度: 今回の評価実験のタスク設定は，SL の語彙概念を検索クエリ，PWN の各 synset を文書，PWN の synset 集合を検索対象の文書集合とみなせば，情報検索のタスクとみなすことができる．特に復元タスクは，情報検索における既知項目検索 (known item retrieval) に対応するので，以下の指標を本実験における性能に関する評価尺度として援用する．

[4]PWN の全 synset の中から頻度の高い 4,997 の synset を抽出したもの．http://wordnetcode.princeton.edu/stand-off-files/core-wordnet.txt

- n 位における成功率 S@n: ここで n は結果におけるランクであり，S@n は，指定されたランク n までに正解が得られるクエリの割合を表す．
- 平均逆順位 MRR (Mean Reciprocal Rank): 正解が得られる最上位のランクの逆数の平均である．

評価における観点: 今回の実験では，以下の観点をパラメータとして走行実験を行った．

- 式 (1) における β を変化させることにより，性能に対する $score_p$，$score_g$ の寄与を調べた．
- SL の gloss テキストを翻訳させる際に，τ_A, τ_B, τ_C を用いた場合を比較する．ここで，τ_C は翻訳に冗長性を持たせることの効果を調べるために設定した．翻訳の冗長構成に関しては，さらに式 (5) による $score'_g(s,t)$ も評価対象とした．

$$score'_g(s,t) = max(TextSim(\tau_A(gloss(s)), gloss(t)), TextSim(\tau_B(gloss(s)), gloss(t))) \quad (5)$$

- 集合間の類似性測度: $TextSim(U,V)$ の具体的な関数形として，代表的な測度である Dice 係数，Simpson 係数を用いた場合を比較した．

4.2 WN-Ja からの復元タスク

全般的な傾向として，$TextSim(U,V)$ で利用する集合間の類似性測度として Dice 係数を用いた場合の結果が一番良好であったため，以下ではこの場合に限って議論する．また紙面の制約から，S@1 および MRR の結果についてのみ議論する．

図 1 は，β を変化させた時の S@1 の値を gloss テキストの翻訳手段ごとにプロットしたものである．ただし，Non-MT はクエリの WN-Ja synset に対応する PWN synset の英語 gloss を用いた場合，すなわち，理想的な翻訳が行われた仮想的な状況を示す．また，τ_M とは，式 (5) を用いた場合の結果である．

この結果から，以下のようなことが言える．

- Non-MT の場合の値は β とともに単調増加するが，その立ち上がりは早く，$\beta = 0.2$ あたりでほぼ飽和に達する．
- 全ての翻訳方法について，gloss テキストの類似性のみを用いる場合 ($\beta = 1.0$) を超えるケースは少なく，またその差は僅か (τ_C の $\beta = 0.08$ において 0.726 に対し，$\beta = 1.0$ の 0.721 など) である．
- τ_C，τ_M の場合の値が，τ_A，τ_B の場合の値を上回っていることから，翻訳機能の冗長構成は有効である．

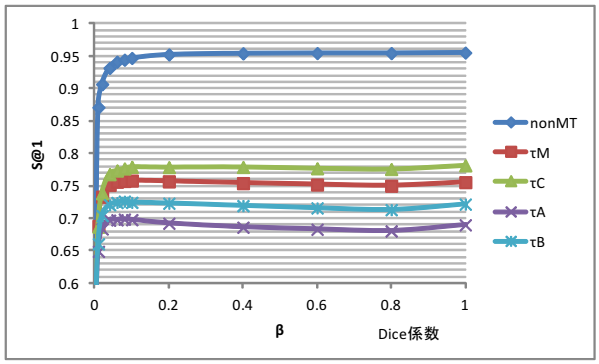

図 1: 復元タスク: S@1 の結果

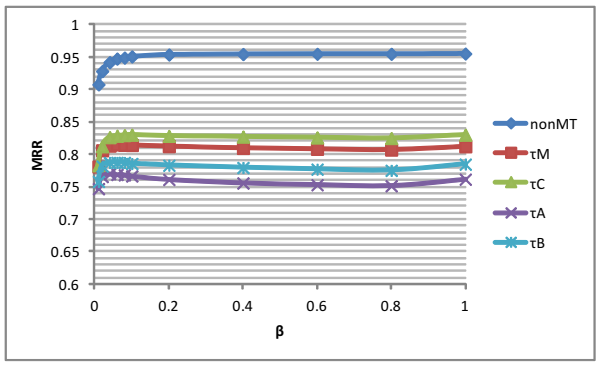

図 2: 復元タスク: MRR の結果

図 2 は，β を変化させた時の MRR の値を同様にプロットしたものである．定義から容易に想像されるように，β の立ち上がりの領域を除けばほぼ S@1 の場合と同様の傾向を示している．

4.3 EDR からの発見タスク

WN-Ja からの復元タスクの場合とは異なり，全般的な傾向として $TextSim(U,V)$ で利用する集合間の類似性測度として Simpson 係数を用いた場合の結果が一番良好であったため，以下ではこの場合に限って議論する．また紙面の制約から，もっとも厳しい尺度である Syn-level の S@1 ともっとも緩やかな尺度である Rel-level の S@10 の結果についてのみ議論する．

図 3 は，β を変化させた時の Syn-level:S@1 の値をプロットしたものである．ただし，Non-MT はクエリの EDR 擬似 synset における英語の gloss テキストを用いた[5]時の値である．

この結果から，以下のようなことが言える．

- $\beta \leq 0.1$ の領域に最大値が現れているが，それ以上 β を大きくしても結果は悪くなる．

[5]EDR の概念には，日本語・英語両方による gloss テキストが与えられているものがあり，今回のクエリセットはこのような EDR 概念を選択している

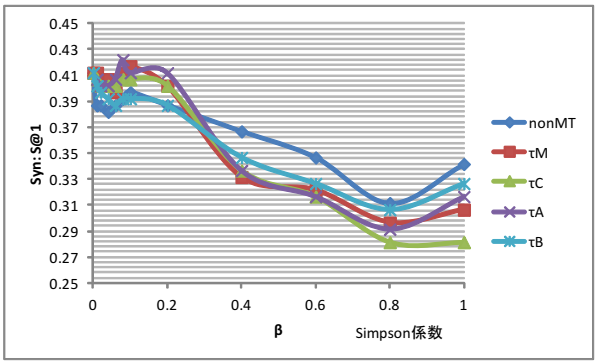

図 3: 発見タスク: Syn-level:S@1 の結果

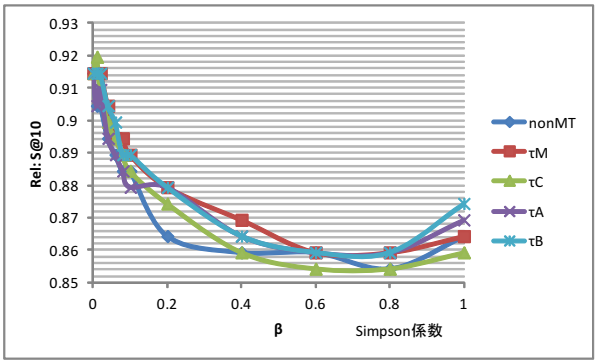

図 4: 発見タスク: Rel-level:S@10 の結果

- 興味深いことに，この領域においては EDR のオリジナルの英語 gloss を用いるよりも日本語 gloss を機械翻訳した方が結果が良好である．
- 復元タスクの場合ほどには翻訳機能の冗長構成の有効性は明確ではない．

図 4 は，β を変化させた時の Rel-level:S@10 の値をプロットしたものである．この評価尺度に対しては，ほぼ全ての翻訳手段において $\beta = 0$ の時に最大値が得られており， $score_g$ の有効性は認められない．また，図 3 の場合と同様に，翻訳機能の冗長構成の有効性は明確ではない．

4.4 まとめと考察

復元タスクにおいて gloss テキスト類似性が極めて有用な手がかりであることは予想どおりであった．一方，発見タスクの結果からは，機械翻訳された gloss テキストの類似性が有用な手がかりとなりうるという示唆が得られたが，評価尺度にわたって安定したスコア統合重みの値を定めることは，相当に困難であることも分かった．これから示唆されることは，機械翻訳された gloss テキストの類似性が有用であるクエリと全くそうでないクエリが存在したことである．今後は，これらのクエリごとの詳細な調査を進め，その判別の手がかりを探索する．また，発見タスクにおける翻訳機能の冗長性による効果も明確ではなかったが，テキスト類似性の測度として Simpson 係数がよりロバストであることが分かった．今後は，用いる冗長性をさらに増やした実験を行うと同時に，gloss テキストを語集合へ変換する際の正規化処理の改良を試みる．

5 おわりに

異言語の語彙概念間の対応付けのための手がかりとして，語彙概念を規定する同義語集合間の対応関係だけでなく，翻訳された gloss テキストの類似性を用いることの有効性について評価した．その結果，有効な手がかりとして利用できる可能性は確認できたものの，適切なスコア統合重みを見出すことは自明なタスクではないことも分かった．今後は，より良い手がかり情報の統合方法を探索するとともに，語彙概念体系における局所的な構造の類似性を考慮する方法を検討する．

謝辞

本研究は，科研費 (21520401) ，総務省戦略的情報通信研究開発推進制度 (SCOPE) の援助を受けた．

参考文献

[1] Satanjeev Banerjee and Ted Pedersen. 2003. Extended Gloss Overlaps as a Measure of Semantic Relatedness. *Proc. IJCAI 2003*, pp.805–810.

[2] Francis Bond, Hitoshi Isahara, et al. 2009. Enhancing the Japanese WordNet. *Proc. the 7th Workshop on Asian Language Resources.*

[3] Jérôme Euzenat and Pavel Shvaiko. 2007. *Ontology Matching.* Springer.

[4] Yoshihiko Hayashi. 2012. Computing Cross-Lingual Synonym Set Similarity by Using Princeton Annotated Gloss Corpus. *Proc. 6th International Global WordNet Conference*, pp.134–141.

[5] 林 良彦. 2012. Princeton Annotated Gloss Corpus を用いた異言語の語彙概念の対応付け. 電子情報通信学会 思考と言語研究会, TL2011-59, pp.29–34.

[6] Michael Lesk. 1986. Automatic Sense Disambiguation Using Machine Readable Dictionaries: How to tell a pine cone from an ice cream cone. *Proc. SIGDOC 86*, pp.24–26.

[7] George A. Miller and Christiane Fellbaum. 2007. WordNet Then and Now. *Language Resources and Evaluation*, Vol.41, pp.209–214.

[8] Toshio Yokoi. 1995. The EDR Electronic Dictionary. *Communications of the ACM*, Volume 38, Issue 11, pp. 42–44.