言語横断関連報道記事検索における翻訳ソフト・対訳辞書・数値表現翻訳規則の性能比較*

浜本 武　中山 健明　日野 浩平　堀内 貴司
豊橋技術科学大学 工学部 情報工学系
{hamamo,takeaki,hino,takashi}@cl.ics.tut.ac.jp

宇津呂 武仁
京都大学大学院 情報学研究科
utsuro@i.kyoto-u.ac.jp

1 はじめに

近年，WWW上の日本国内の新聞社などのサイトにおいては，日本語だけでなく英語で書かれた報道記事も掲載しており，これらの英語記事においては，同一時期の日本語記事とほぼ同じ内容の報道が含まれている．これらの日本語および英語の報道記事のページにおいては，最新の情報が日々刻々と更新されており，分野特有の新出語（造語）や言い回しなどの翻訳知識を得るための情報源として，非常に有用である．本研究では，これらの報道記事のページから日本語および英語など，異なった言語で書かれた文書を収集し，多種多様な分野について，分野固有の固有名詞(固有表現)や事象・言い回しなどの翻訳知識を自動または半自動で獲得する手法についての研究を行う．

本研究における日英関連報道記事からの翻訳知識獲得の流れを図1に示す[堀内02, Utsuro02]．まず，翻訳知識獲得のための情報源収集を目的として，同時期に日英二言語で書かれたWWW上の新聞社やテレビ局のサイトから，報道内容がほぼ同一もしくは密接に関連した日本語記事および英語記事を検索する．そして，取得された関連記事対に対し，内容的に対応する翻訳部分の推定を行い，その推定範囲から二言語間の訳語対応を推定し，訳語対の獲得を行う[堀内03, 日野03]．

この一連の枠組において，特に本論文では，言語横断報道記事検索過程に焦点をあて，英語記事と日本語記事の間で，言語を横断して記事の類似性の度合を推定する情報源の性能を評価する．具体的には，翻訳ソフト，対訳辞書，および，記事中の数値表現の二言語間対応付けを行うための数値表現翻訳規則の三種類をとりあげ，これらの各種情報源の性能を比較する．それぞれの情報源の特徴として，翻訳ソフトでは，翻訳結果が一意に決定されるため，関連記事とは無関係な記事が検索される可能性は少ないが，関連記事自体を見落とす可能性がある．逆に対訳辞書では，ある単語に対する様々な訳語が提供されるために関連記事が検索される割合は高くなるが，無関係な記事も多く検索されてしまう．また，報道内容がほぼ同一の二言語記事間である場合，同一の数値表現が現れる可能性が高いため，数値表現が多く含まれる記事の数値表現を適切に翻訳できれば，ほぼ同一の内容が書かれた相手言語記事を検索できる可能性がある．評価実験においては，まず，各々の情報源をそれぞれ単独で利用した場合の検索性能を評価した後，これらの情報源を統合した場合の検索性能を評価する．

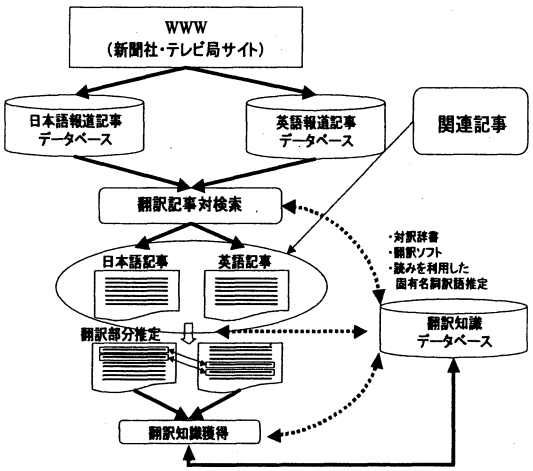

図1: 日英関連報道記事からの翻訳知識獲得の流れ

2 言語横断関連報道記事検索

言語横断関連報道記事検索の流れを図2に示す．

まず，新聞社やテレビ局のサイトから英語記事 d_E と日本語記事 d_J を取得する．次に，関連記事対はお互いの日付が近いと想定して，日付の情報を用いて検索対象の記事を絞りこむ．そして，取得した英語記事 d_E と日本語記事 d_J の間の類似性を測るために，翻訳ソフト・対訳辞書・数値表現翻訳規則などの情報源を利用して英語記事 d_E を日本語訳 d_{trJ} に変換し，この日本語訳 d_{trJ} と日本語記事 d_J から翻訳頻度ベクトル $v(d_{trJ})$ と日本語頻度ベクトル $v(d_J)$ をそれぞれ作成する．最後に，頻度ベクトル間で類似度を計算し[1]，類

*Performance Comparison of MT Software, Bilingual Lexicon, and Translation Rules of Numerical Expressions in Cross-Language Retrieval of Relevant News Articles

[1] 平仮名語の高頻度機能的表現26語を不要語として削除した．ま

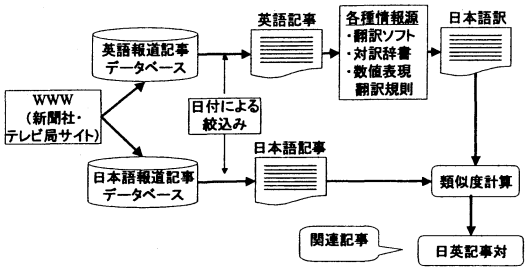

図 2: 日英関連報道記事検索の流れ

似度が下限値以上の記事を検索結果とする.

2.1 翻訳ソフト

翻訳ソフトを用いた記事間類似度の測定においては, まず, 市販の翻訳ソフトを用いて英語記事 d_E を日本語訳 d_{trJ}^{MT} に翻訳する. 次に, 日本語訳 d_{trJ}^{MT} と日本語記事 d_J を, 日本語形態素解析システム「茶筌」によって形態素解析し, 形態素の頻度ベクトル $v(d_{trJ}^{MT})$ および $v(d_J)$ をそれぞれ作成する. そして, 翻訳ソフトを用いた場合の記事間類似度 $sim_{MT}(d_E, d_J)$ としては, これらの頻度ベクトルの余弦を用いる.

2.2 対訳辞書

対訳辞書を用いて記事間類似度を測定する場合は, まず, 英語記事 d_E を空白毎に分割して得られる各英語単語 $w_E^i (i = 1, \ldots, N_E)$ に対して, 既存の対訳辞書 (英辞郎 Ver.37: 見出し語 85 万語) を利用して英語単語の日本語訳集合 W_{trJ}^i を得る. 次に, 英語記事 d_E の各英語単語における日本語訳集合 W_{trJ}^i の和集合を英語記事の日本語訳 d_{trJ}^{BL} とみなす.

$$d_{trJ}^{BL} = \bigcup_{i=1,\ldots,N_E} W_{trJ}^i$$

そして, 日本語形態素解析システム「茶筌」を用いてこの日本語訳 d_{trJ}^{BL} と日本語記事 d_J を形態素解析し, 形態素の頻度ベクトル $v(d_{trJ}^{BL}), v(d_J)$ をそれぞれ作成する. 対訳辞書を用いた場合の記事間類似度 $sim_{BL}(d_E, d_J)$ としては, これらの頻度ベクトルの余弦を用いる.

2.3 数値表現翻訳規則

数値表現翻訳規則を用いて記事間類似度を測定する場合は, まず, 数値表現抽出規則[2] を用いることにより, 英語記事 d_E および日本語記事 d_J から, 数値表現集合 d_E^{NM} および d_J^{NM} をそれぞれ作成する. さらに, 数値表現翻訳規則[3] を用いて英語数値表現集合 d_E^{NM} を日本語訳数値表現集合 d_{trJ}^{NM} に翻訳し, これらの数値表現集合から, 頻度ベクトル $v(d_{trJ}^{NM})$ および $v(d_J^{NM})$ をそれぞれ作成する. 次に, これらの頻度ベクトル間で, 以下の重み付き一致数を求め, これを $CorrNum(v(d_{trJ}^{NM}), v(d_J^{NM}))$ とする.

$$\sum_{i=1}^{n} w_{nm} \cdot \min\left\{v_i(d_{trJ}^{NM}), v_i(d_J^{NM})\right\}$$

$$w_{nm} = \begin{cases} 1 & \text{(数値部と単位部から構成される数値表現)} \\ 0.5 & \text{(数値部のみから構成される数値表現)} \end{cases}$$

最後に, 評価対象となっている全検索質問記事におけるこの重み付き一致数 $CorrNum$ の最大値を $MaxCorrNum$ として, $MaxCorrNum$ で各重み付き一致数 $CorrNum(v(d_{trJ}^{NM}), v(d_J^{NM}))$ を正規化した値

$$\frac{CorrNum(v(d_{trJ}^{NM}), v(d_J^{NM}))}{MaxCorrNum}$$

を数値表現翻訳規則を用いた場合の記事間類似度 $sim_{NM}(d_E, d_J)$ とする. 数値表現翻訳規則を用いた記事間類似度計算の例を図 3 に示す.

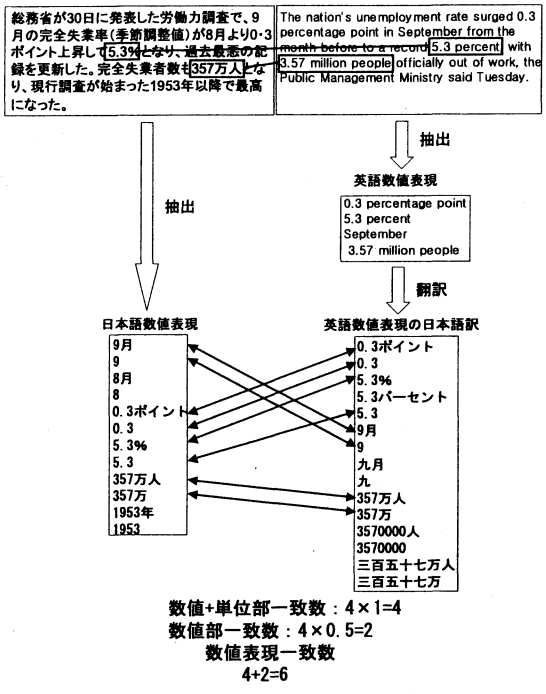

図 3: 数値表現翻訳規則を用いた記事間類似度計算の例

[2] 数値表現は, 一般に, 数値情報 (以下, 数値部), および, 数値に付与されている単位 (以下, 単位部) から構成される. 英語数値表現抽出規則の作成においては, まず, 187 個の数値部抽出規則, 24 個の前接単位部規則, および, 110 個の後接単位部規則をそれぞれ人手で作成し, これらの任意の組合せを英語数値表現抽出規則とした. また, 日本語数値表現抽出規則は, 日本語形態素解析システム「茶筌」の品詞分類を用いて, 正規表現パターン "(数詞+ 名詞*)+" として記述した.

[3] 数値部に関する 4 つの数値表現翻訳規則 (整数, 小数, 分数, 時間表記) と単位部に関する 433 の数値表現翻訳規則を人手で作成し, これを数値表現翻訳規則とした.

表 1: 平均記事数・平均記事長・日付のずれ

サイト	一日の平均記事数		一記事の平均記事長 (byte)		日付のずれ (日)
	英語	日本語	英語	日本語	
A	1.1	36.9	1087.3	759.9	± 4
B	18.0	88.4	3135.5	836.4	± 3
C	21.2	97.4	3228.9	837.7	± 2

2.4 複数情報源の統合

翻訳ソフト，対訳辞書，数値表現翻訳規則を用いた場合の記事間類似度の重みをそれぞれ w_{MT}，w_{BL}，w_{NM} とすると，三種類の情報源を統合した場合の記事間類似度 sim_{MIX} は，以下の重み付き和で定義される．

$$sim_{MIX} = w_{MT} \cdot sim_{MT} + w_{BL} \cdot sim_{BL} + w_{NM} \cdot sim_{NM}$$
$$0 \leq w_{MT}, w_{BL}, w_{NM} \leq 1 \quad , \quad w_{MT} + w_{BL} + w_{NM} = 1$$

3 実験および評価

本論文の評価実験では，A~C の三種類のサイトから収集した日本語および英語の報道記事を用いた．各サイトにおける一日の平均記事数，一記事の平均記事長，および，相手言語においてほぼ同一の内容の記事が存在する場合の日付のずれを表 1 に示す．一日の平均記事数については，三サイトとも英語記事よりも日本語記事の方が多い．また，これらのサイトにおいては，表に示した日付の範囲であれば，英語記事から日本語記事を検索する方向で言語横断関連報道記事収集を行えば，5 割以上の率で，ほぼ同じ内容の日英記事対が収集できる [堀内 02]．

3.1 言語横断関連報道記事検索

各サイトについて，評価用関連記事組を 50 組ずつ収集し，評価用記事対の英語記事を検索質問として，評価用記事対の日本語記事を含む記事集合に対して言語を横断した記事検索を行い，記事間類似度の下限の条件を満たす日本語記事を検索した場合の適合率・再現率の変化をプロットした．この場合の適合率・再現率の定義は，日付の範囲内の評価用記事対の集合を DP_{ref}，記事間類似度を $sim(d_E, d_J)$，記事間類似度の下限値を L_d として，

$$\text{適合率} = \frac{|\{d_J \mid \exists d_E, \langle d_E, d_J \rangle \in DP_{ref}, sim(d_E, d_J) \geq L_d\}|}{|\{d_J \mid \exists d_E \exists d'_J, \langle d_E, d'_J \rangle \in DP_{ref}, sim(d_E, d_J) \geq L_d\}|}$$
$$\text{再現率} = \frac{|\{d_J \mid \exists d_E, \langle d_E, d_J \rangle \in DP_{ref}, sim(d_E, d_J) \geq L_d\}|}{|\{d_J \mid \exists d_E, \langle d_E, d_J \rangle \in DP_{ref}\}|}$$

となる．

単独の情報源を用いて言語横断検索を行った場合の性能を比較したものを図 4 に示す．単独の情報源では，翻訳ソフトの性能が最も高い．図 5~ 図 7 では，翻訳ソフト単独，二種情報源統合，および，三種情報源統合の性能を比較して示す．ただし，各種情報源の重み w_{MT}，w_{BL}，および，w_{NM} に関しては，0.1 刻みで，言語横断関連記事検索の性能が最も高くなる設定を探索し，最適な重みでの結果を示す．全体の傾向としては，二種類の情報源を統合することにより，翻訳ソフト単独の性能が改善でき，さらに，三種類の情報源を統合することにより，ほぼ最大の性能が達成できている．ただし，図 6 および図 7 から分かるように，一部の例外を除いて，数値表現翻訳規則と他の情報源の二種類の混合の段階で，三種類の情報源の統合に匹敵する性能が達成できている．このことから，数値表現翻訳規則は，単独では性能が低いものの，他の情報源と相補的に用いることにより，高い性能を示す特性を持つと言える．

3.2 日英関連報道記事からの訳語対応の推定

次に，[堀内 03] の枠組において，言語横断関連記事検索により関連記事組を収集した結果から訳語対応を推定するタスクにおいて，関連記事検索の性能の改善と訳語対応推定の性能の相関を評価した．言語横断関連記事検索においては，前節で述べたように，複数の情報源を統合することにより，関連記事検索の性能が改善された．一方，訳語対応推定の性能においても，関連記事検索における複数の情報源を統合することにより，全般的に，性能を改善する傾向が観測された．今後は，言語横断関連記事検索におけるどのような特性が，訳語対応推定の性能と相関を持つのかについての分析を詳細に行う必要がある．

4 おわりに

本論文では，日英関連報道記事からの翻訳知識獲得の枠組において，言語横断報道記事検索過程に焦点をあて，言語を横断して記事の類似性の度合を推定する情報源として，翻訳ソフト，対訳辞書，数値表現翻訳規則の三種類の性能を比較した．単独の情報源を利用する場合は翻訳ソフトの性能が最も高かった．また，三種類の情報源を統合することにより，単独の情報源よりも高い検索性能が達成できた．

参考文献

[日野 03] 日野浩平, 堀内貴司, 浜本武, 中山健明, 宇津呂武仁 : 日英関連報道記事からの翻訳知識獲得のためのユーザインタフェースの作成, 言語処理学会第 9 回年次大会論文集 (2003).

[堀内 02] 堀内貴司, 千葉靖伸, 浜本武, 宇津呂武仁 : 言語横断検索により自動収集された日英関連報道記事からの訳語対応の獲得, 情報処理学会研究報告, Vol. 2002, No. (2002-NL-150), pp. 191-198 (2002).

[堀内 03] 堀内貴司, 日野浩平, 浜本武, 中山健明, 宇津呂武仁 : 日英報道記事からの訳語対獲得における言語横断情報検索の有効性の評価, 言語処理学会第 9 回年次大会論文集 (2003).

[Utsuro02] Utsuro, T., et al.: Semi-automatic Compilation of Bilingual Lexicon Entries from Cross-Lingually Relevant News Articles on WWW News Sites, *Machine Translation: From Research to Real Users*, Lecture Notes in Artificial Intelligence: Vol. 2499, pp. 165-176, Springer (2002).

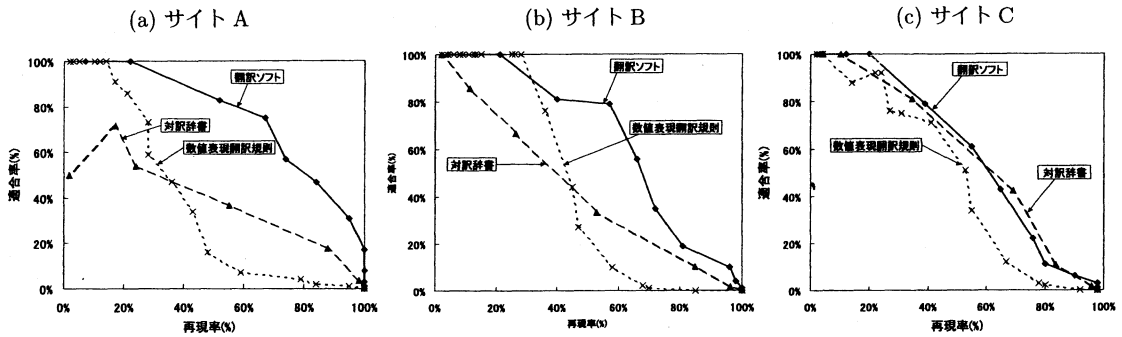

図 4: 単独の情報源を用いた場合の言語横断関連報道記事検索の性能

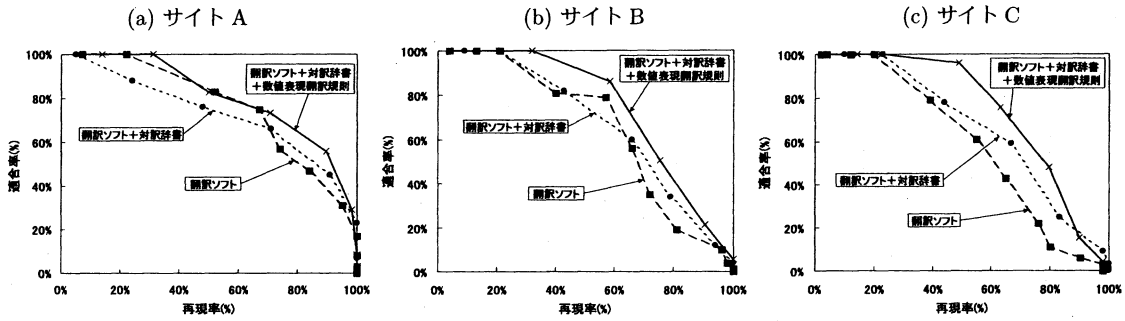

図 5: 翻訳ソフト・翻訳ソフト+対訳辞書・三種情報源統合の性能比較

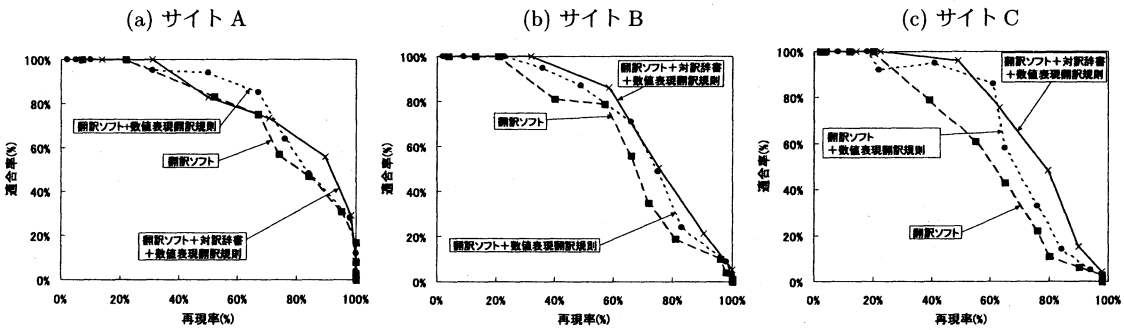

図 6: 翻訳ソフト・翻訳ソフト+数値表現翻訳規則・三種情報源統合の性能比較

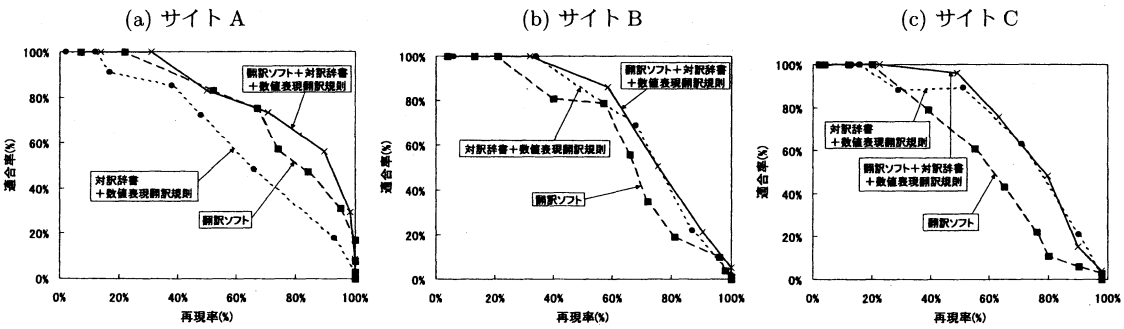

図 7: 翻訳ソフト・対訳辞書+数値表現翻訳規則・三種情報源統合の性能比較