TR-I-0320

用例主導型機械翻訳における情報量による類似検索手法の検討

"A Study on the Effective Measure by Shannon's Information Value in Example-based Machine Translation"

Takashi Okada

1993.3

概要

用例主導型機械翻訳において、名詞句「N1のN2」の場合、品詞・類語など7種類の属性の値の一致性から類似度を計算するのだが、この時、最良の類似度を持つ用例が多く出現することがある。そこで今回、同じ類似度のため優劣のつかない用例に対する効果的な類似選択手法の検討とその実験を行った。本項では、各属性値の持つ情報量の一致性を基準とした類似選択方式を検討した。その結果、この方式を用例主導型機械翻訳システムの類似用例検索処理に組み込むことにより、従来より品質の高い類似用例が選択でき、翻訳の質が向上した。

ATR自動翻訳電話研究所

ATR Interpreting Telephony Research Laboratories

1．はじめに

文献［1］の用例主導型機械翻訳（ＥＢＭＴ）では、入力文と類似な用例（原言語表現と目的言語表現の対）を、シソーラス上の類語コードや品詞などを対象とした類似度計算により検索し、それを利用して翻訳を行う。この時、最良の類似度を持つ用例が多く出現することがあり、これらの用例からどの用例を選択するかが翻訳結果に影響を与える。特に、名詞句「N1のN2」（N1，N2は名詞句）のＥＢＭＴシテスムでは、品詞・類語など7種類の属性値の一致性から類似度を算出するが、しばしば、同じ類似度の用例が多く発生する。今回、この問題を検討するが、その目的・方法・手順を以下に示す。

【目的】最良の類似度を持つ複数の用例の中からより類似な用例を選択する尺度を求める。

【アプローチ方法】類似度計算に使用した属性について、属性値の持つ情報量（シャノンの情報量）の翻訳正解率に対する効果を調べ、その効果を利用して、最良の類似度を持つ用例に対する選択尺度を求める。

【手順】用例データベースは、「国際会議の登録」に関する対話の日英対訳データベースから２５５０件の用例（日英対訳データ）を抽出したもので、この用例データベースから、１用例づつ抜き出してそれを入力とするジャックナイフ・テストを、以下の実験で使用した。

[1] ２５５０件中の最良類似度を持つ複数の用例（２００３件）の翻訳正解率に対するシャノンの自己情報量の特徴を明らかにする。
[2] ２００３件に対し、求められた特徴に基づき、最良の類似度を持つ用例内から最適な用例を選択する尺度を求める。
[3] 求められた尺度をＥＢＭＴシステムに組み込み、類似用例５件を抽出する比較実験を行い、翻訳正解率の向上効果を調べた。

以下で、その説明をする。

2．類似度計算の概略

ＥＢＭＴシステムは、類似用例検索処理のため、用例データベースの他にシソーラス・データベースを持つ。類似用例検索処理は、入力と用例との一致検索に失敗した場合に実行される処理で、品詞などの他に3階層のシソーラス上での類語コードを属性として扱う類似度計算（図１）を、全用例に対して行い、類似用例を選択する処理である。また、入力と用例との類語間の距離の見積り方法を図２に示す。

７種類の属性（「N1」・「N2」の品詞・接辞・類語コード、「の」の文字列パターン）から、以下のようにして類似度を計算する。（類似度が小さいほど類似性が高い。）

$$\text{類似度 } D(I, E) := \sum_{i=1}^{7} w(I_i) \cdot d(I_i, E_i)$$

∴ Ｉ：入力　Ｅ：用例

∴ $w(I_i)$：入力のｉ番目の属性の属性値の日英変換における相違からくる重み

∴ $d(I_i, E_i)$：入力と用例のｉ番目の属性の属性値間の距離
（類語のみ、０，１／３，２／３，１、その他は、０，１）

図１．類似度計算

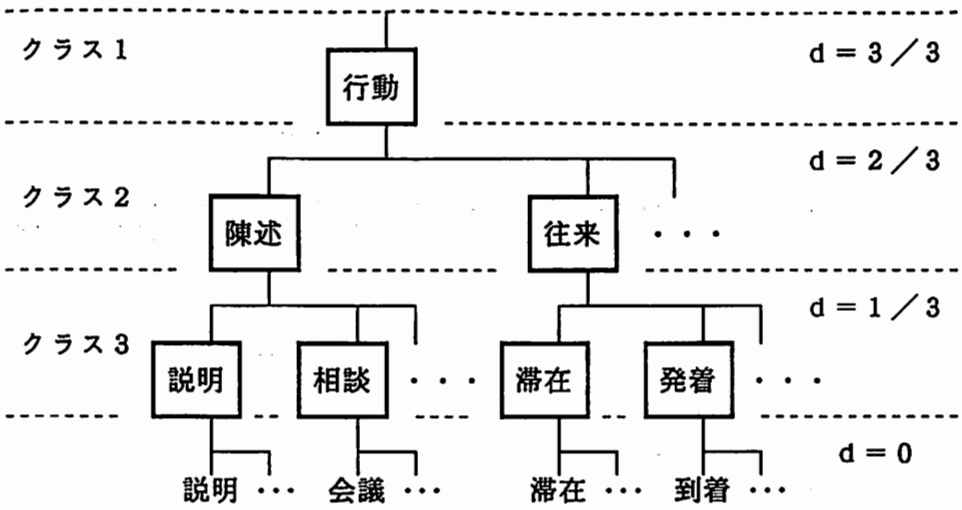


図2．シソーラスの階層と類語コードの距離（d）

表1．２５５０件の用例の属性とその値（属性値）の分布について

	N1			「の」	N2		
	品詞	類語	接辞		品詞	類語	接辞
属性値の総数	2550	4136	2550	2550	2550	4383	2550
種類	9/12	大： 12／ 12 中： 80／102 小：275/1002	1／3	1／8	7/12	大： 11／ 12 中： 81／102 小：287/1002	1／3
平均情報量 $-\Sigma p_i \cdot \log_2 p_i$	1.87	大：2.79388 中：4.68875 小：6.21088	0.0	0.0	1.03	大：2.65 中：4.44 小：5.68	0.0
異なる属性値間の距離	1 or 0	大：1 中：2/3 小：1/3 or 0	1 or 0	1 or 0	1 or 0	大：1 中：2/3 小：1/3 or 0	1 or 0
重み <最大値>	1.0	1.0	0.46	0.46	1.0	1.0	0.46
重み <最小値>	0.45	0.43598	0.46	0.46	0.48	0.40	0.46
重み <平均値>	0.52	0.71878	0.46	0.46	0.53	0.79	0.46

注）類語の大分類が１０種類（’０’～’９’）より多いのは、特殊コードとして”ｘｘｘ”（未登録語），”ｙｙｙ”（専門用語）の２種類を加わっているからである。距離計算上、「N1」「N2」が両方”ｙｙｙ”の時のみ距離０、他の場合（１つでも特殊コードを含んでいる場合）、距離１とするルールを持つ。

注）類語数が用例数よりも多い（「N1」の場合、４１３６個）のは、１つの単語が複数の類語の意味を持つためである。距離計算時は、単語間で成り立つ全ての類語対の内、最も距離の小さいものが選択される。

注）「の」の助詞は、８種類：「の」・「への」・「との」・「からの」・「よりの」・「での」・「までの」・「についての」

注）接辞は、３種類：接頭語・接尾語・接辞なし

注）品詞は、１２種類：普通名詞・固有名詞・サ変名詞・代名詞・副詞・形容詞・副助詞・記号・数詞・連体詞・格助詞・形容名詞

用例データベースは、「国際会議の登録」に関する対話の日英対訳データベースから２５５０件の用例（日英対訳データ）を抽出したもので、２５５０件の用例の各種属性値の分布について、表１に示す。

３．同じ類似度を持つ用例の出現状況

２５５０件の用例からなる用例データベースから１用例づつ抜き出してそれを入力とするジャックナイフ・テストを２５５０回繰り返した。その結果、同じ類似度を持つ多くの用例が類似検索で出現していることが分かった（表２）。ちなみに、最小の類似度を持つ（最も類似な）用例数は、平均２７．２個で、それは１～１１９０個の範囲に分散し、類似度において、小さい方から２番目以降も同様に、同じ類似度を持つ多くの用例が出現している。

表２．同じ類似度を持つ用例の用例数一覧

Ｎｏ	事例内の類似度の順序	平均用例数	平均類似度
1	最小の類似度	27.20	0.13
2	２番目に小さい類似度	17.75	0.42
3	３番目に小さい類似度	29.10	0.57
4	４番目に小さい類似度	31.46	0.68

４．シャノンの情報量の利用方法

シャノンの（統計的）情報量は、事象の生ずる確率ｐ（≦１）を情報量Ｉ（ｐ）として、

$$I(p) = \log_2 \frac{1}{p} = -\log_2 p \quad (\text{ビット})$$

と定義される。（文献［２］参照）

名詞句「Ｎ1のＮ2」の２５５０件の用例から、各属性の属性値に対して、頻度に基づく情報量を算出する。また、入力と用例の属性値の一致／不一致から各属性の属性値ごとに一致／不一致成分の情報量を算出する（図３）。

「類語コード」以外の属性については、一方の値が０、他方が属性値の情報量になるが、「類語コード」の場合、３段階の内の一致レベルのコードの持つ情報量を一致成分の情報量とするため、一致／不一致成分の情報量が共に０にならない場合がある。その例として、入力の「Ｎ2」が『会議』、用例が『滞在』の場合を表３に示し、それらのシソーラス上の情報量の分布を図４に示す。

属性値の持つ情報量：２５５０件の用例内の頻度ｐに基づく情報量
一致成分の情報量　：属性値内の一致情報の持つ情報量
不一致成分の情報量：属性値の持つ情報量 － 一致成分の情報量

図３．属性値の情報量（全体・一致／不一致成分）の算出方法

表３．「N2」について、入力が『会議』、用例が『滞在』の場合の情報量の例

「N2」		品詞	類語コード	情報量	品詞	類語
『会議』	属性値	サ変名	"344":行動->陳述->相談	一致	3.14	4.42
	情報量	3.14	5.47:4.42->5.14->5.47	不一致	0.0	1.05
『滞在』	属性値	サ変名	"319":行動->往来->滞在	一致	3.14	4.42
	情報量	3.14	12.10:4.42->8.78->12.1	不一致	0.0	7.68

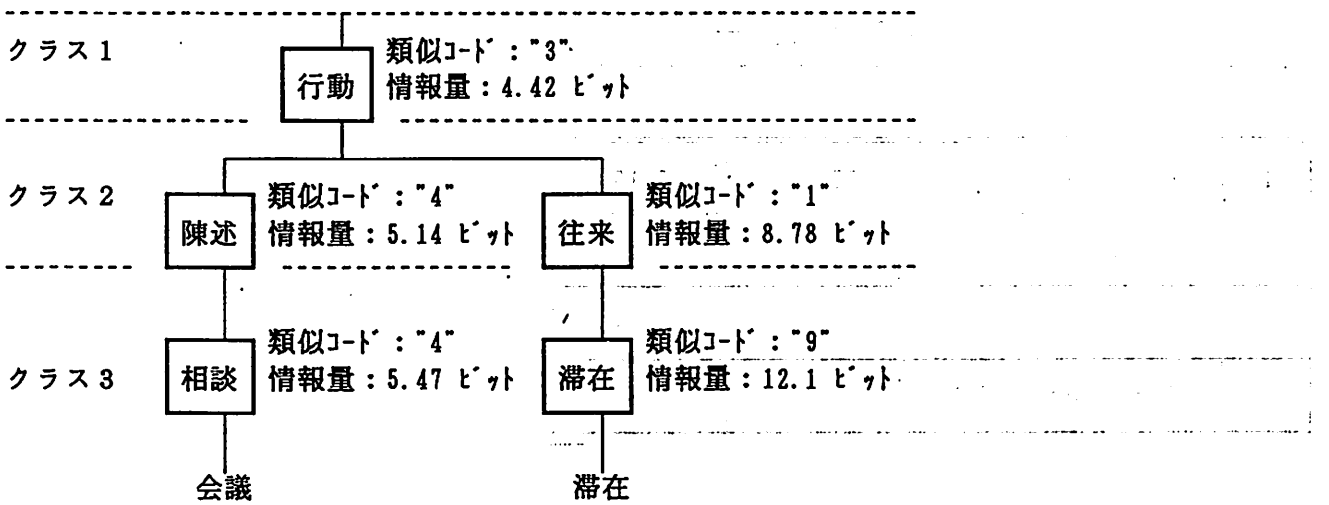


図４．『会議』・『滞在』のシソーラス上の情報量の分布

5．翻訳正解率に対する情報量の特徴

２５５０件の検索事例の内、最良の類似度を持つ用例を複数個持つ事例２００３件を対象に、情報量の大小や近似度で並べ、翻訳正解率の分布を調べる実験をした。翻訳の正解／不正解は、類似用例の翻訳パターンが入力に使用した用例の翻訳パターンと一致しているか否かを基準にした。

この２００３件の事例について、最良の類似度を持つ用例の品質を表３に示す。

表３．２００３件の事例の品質一覧

事例の分類	事例数	平均類似度	平均用例数	翻訳正解率の平均	翻訳正解率の上限
事例全体	2003	0.12	34.4	74.3 %	87 %
類似度０の事例	1514	0.0	37.7	86.5 %	95 %
類似度０以外の事例	489	0.48	24.0	36.8 %	63 %

また、以下で使用する記号の説明をする。

i_j　：入力『N1のN2』のｊ番目の属性の情報量
e_j　：用例『N1のN2』のｊ番目の属性の情報量
s_x　：情報量の演算式の結果の標準偏差値
s_y　：翻訳正解率の標準偏差値
x　：情報量の演算式の結果の値
効果Ｅ：［資料－１］参照

○ 入力と用例との情報量における近接性の実験とその結果

入力と用例の情報量について、属性単位、及び、全属性の総和に基づいて、入力と用例との情報量の近接度と翻訳正解率の関係を実験で求めた。その結果、以下の２種類の効果的な演算式を得た。実験では、同じ類似度を持つ用例を演算式の値で昇順に並べて、翻訳正解率の分布を調べた。

表４．２００３件の全事例での近接性の実験結果

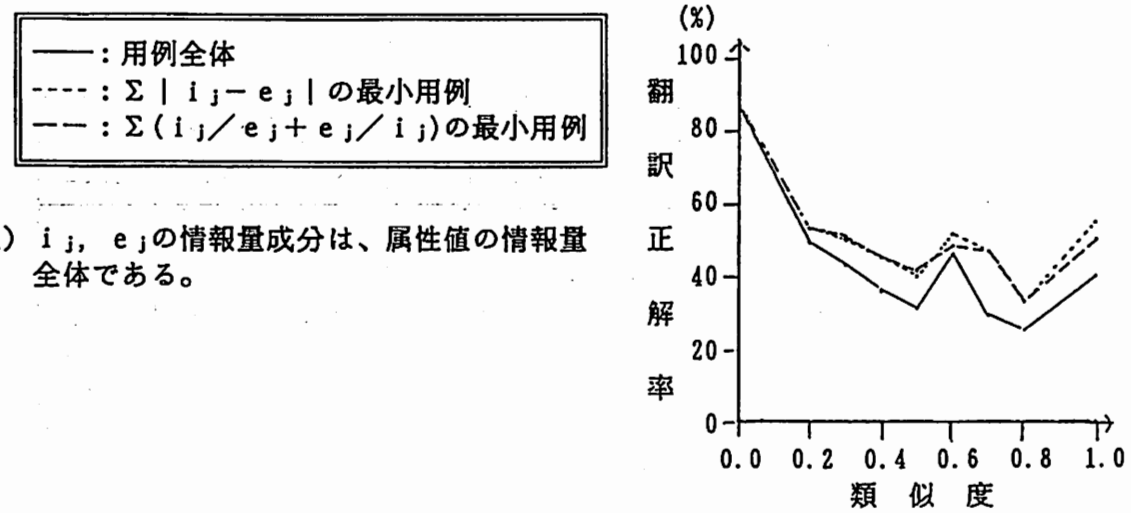
情報量の演算式	i_j, e_jの情報量成分	1 stの正解率	(1 st－last)の正解率	1 stの s_x/x	1 stの s_y	1 stの効果Ｅ
$\sum_{j=1}^{7} \lvert i_j - e_j \rvert$	情報量全体	77.1 %	7.84 %	2.515	0.420	0.327
	一致成分	76.8 %	7.39 %	3.356	0.422	0.201
$\sum_{j=1}^{7} (i_j / e_j + e_j / i_j)$	情報量全体	77.2 %	8.89 %	0.125	0.419	0.149
	一致成分	76.9 %	7.99 %	0.078	0.421	0.195

表５．類似度０（１５１４件）の事例での近接性の実験結果

情報量の演算式	i_j, e_jの情報量成分	1stの正解率	(1st−last)の正解率	s_x/x	s_y	効果E
$\sum_{j=1}^{7} \lvert i_j - e_j \rvert$	情報量全体	87.3 %	5.94 %	5.450	0.333	0.118
	一致成分	同上	同上	同上	同上	同上
$\sum_{j=1}^{7} (i_j/e_j + e_j/i_j)$	情報量全体	87.4 %	6.61 %	0.010	0.332	0.074
	一致成分	同上	同上	同上	同上	同上

表６．類似度０以外（４８９件）の事例での近接性の実験結果

情報量の演算式	i_j, e_jの情報量成分	1stの正解率	(1st−last)の正解率	s_x/x	s_y	効果E
$\sum_{j=1}^{7} \lvert i_j - e_j \rvert$	情報量全体	45.4 %	13.70 %	1.019	0.498	0.097
	一致成分	44.4 %	11.86 %	1.815	0.497	0.041
	不一致成分	43.6 %	12.07 %	1.112	0.496	0.062
$\sum_{j=1}^{7} (i_j/e_j + e_j/i_j)$	情報量全体	45.8 %	15.95 %	0.234	0.499	0.089
	一致成分	44.6 %	12.27 %	0.153	0.498	0.062
	不一致成分	43.6 %	11.04 %	1.335	0.496	0.036



注）i_j, e_jの情報量成分は、属性値の情報量全体である。

図５．類似度に対する翻訳正解率の分布図
（２００３件の全事例を対象とする）

表４・表５・表６から、入力と用例の各情報量の近接性に基づく用例の類似選択が、翻訳正解率を高めることが分かる。

また、図５は、入力と用例の各情報量の近接度に基づく選択用例の類似度に対する翻訳正解率の分布を表したグラフである。このグラフを見ると、これらの選択用例が、類似度にかかわらず常に高い翻訳正解率を示している。また、類似度が高いほど高い翻訳正解率を示す傾向がある。

実験結果として、以下のような情報量における近接性の特徴を得た。

【特徴ａ】同じ類似度を持つ用例内で、各属性の情報量が入力の各属性の情報量と近接している用例ほど、翻訳正解率が高い。

○ 情報量の大きさと翻訳正解率との関係についての実験とその結果

入力と用例の情報量について、情報量を属性値全体・一致成分・不一致成分にわけて、その情報量の大小と翻訳正解率との関係を実験で調べた。その結果、以下の効果的な演算式を得た。実験では、同じ類似度を持つ用例を演算式の値で昇順に並べて、翻訳正解率の分布を調べた。

表７．類似度０以外（４８９件）の事例での情報量の大小の実験結果

情報量演算式	i_j, e_jの情報量成分	1stの正解率	(1st−last)の正解率	1stの s_x/x	1stの s_y	1stの効果Ｅ
$\sum_{j=1}^{7} i_j$	一致成分	45.2 %	14.93 %	0.274	0.498	0.027
	不一致成分	45.8 %	15.75 %	0.451	0.499	0.009

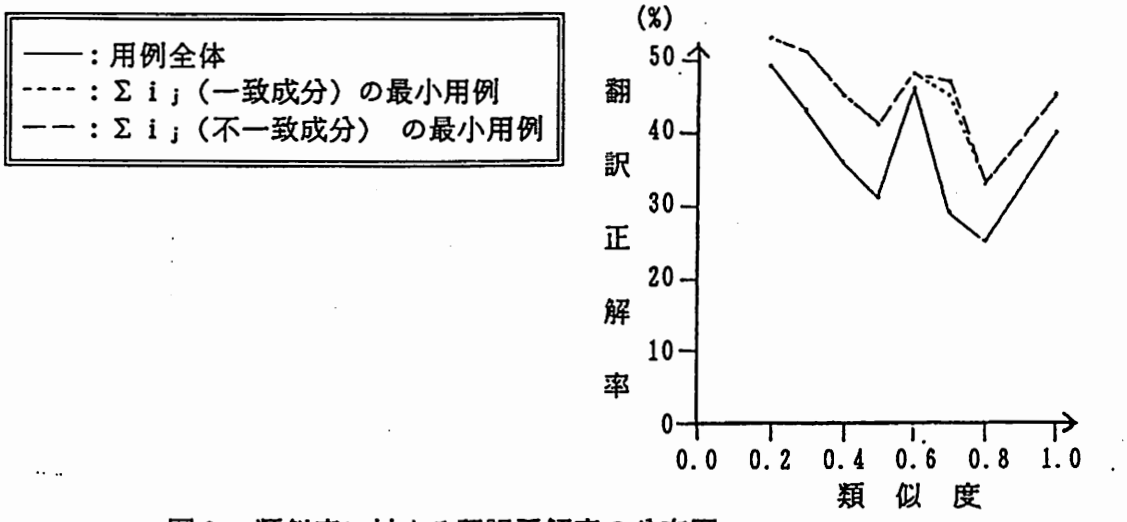


図６．類似度に対する翻訳正解率の分布図
（類似度０以外の事例４８９件を対象とする）

表７から、入力の情報量の一致／不一致成分の総和の小ささに基づく用例の類似選択が、翻訳正解率を高めることが分かる。

また、図６は、入力の情報量の一致／不一致成分の総和の小ささに基づく選択用例の類似度に対する翻訳正解率の分布を表したグラフである。このグラフを見ると、これらの選択用例が、類似度にかかわらず、常に高い翻訳正解率を示している。

実験結果として、以下のような情報量の特徴を得た。

【特徴ｂ】同じ類似度を持つ用例内で、用例に対する入力の一致成分（または、不一致成分）の情報量の総和が小さいほど、翻訳正解率が高い。

6．情報量の特徴に基づく効果的な用例選択基準

次に、上述の実験で求めた情報量の翻訳正解率に対する特徴【特徴ａ】・【特徴ｂ】に基づいた効果的な用例選択基準を求める実験をした。上述の演算式の合成（加算・乗算）を行った結果、図７に示す３種類の尺度が効果的であることが分かった。以下で、それらを簡単に説明する。

［尺度Ａ］　【特徴ａ】に基づくものである。

［尺度Ｂ］　【特徴ａ】に基づくが、情報量を一致成分／不一致成分に分けて演算式を作成し、その合成（加算）により求めた。

［尺度Ｃ］　【特徴ａ】・【特徴ｂ】に基づく。情報量を一致成分／不一致成分に分けて演算式を作成し、その合成（加算・乗算）により求めた。

その結果、表８に示すように、最良の類似度を複数個持つ用例から尺度による値が最も小さい（１st）用例を選択する実験の結果、２００３件の全事例で約３％、類似度０の事例１５１４件で約１％、類似度０以外の事例４８９件で約９％の翻訳正解率の向上が認められた。

尺度Ａ：$\Sigma_{j=1,7} \left(i_j / e_j + e_j / i_j \right)$

尺度Ｂ：$\left(\Sigma_{j=\text{一致成分のみ}} \left(| i_j - e_j | \times \left(i_j / e_j + e_j / i_j \right) \right) \right) / \sigma_{RD} + \left(\Sigma_{j=\text{不一致成分のみ}} \left(i_j / e_j + e_j / i_j \right) \right) / \sigma_R$

尺度Ｃ：$\left(\Sigma_{j=\text{一致成分のみ}} | i_j - e_j | \times \Sigma_{j=\text{一致成分のみ}} i_j \right) / \sigma_{DI} + \left(\Sigma_{j=\text{不一致成分のみ}} \left(i_j / e_j + e_j / i_j \right) \right) / \sigma_R$

注）「i_j」は、入力の属性値の情報量で、「e_j」は、用例の属性値の情報量である。
注）「σ_{RD}, σ_{DI}, σ_R」は、各演算部分の標準偏差の値で、それぞれの分散を１にすることにより測定単位の変化量を標準化した。（文献［３］参照）

図７．情報量に基づく効果的な用例選択基準

表８．三種類の効果的な尺度による実験結果

尺度	事例の分類	１stの正解率	(１st－last)の正解率	s_x/x	s_y	効果Ｅ
尺度Ｂ	全事例（2003件）	77.2 %	8.44 %	8.928	0.419	0.093
	類似度０の事例(1514件)	87.3 %	5.94 %	6.007	0.333	0.112
	類似度０以外の事例(489件)	46.0 %	16.16 %	4.813	0.499	0.026
尺度Ｃ	全事例（2003件）	77.2 %	8.34 %	2.679	0.420	0.183
	類似度０の事例(1514件)	87.3 %	5.94 %	5.532	0.333	0.115
	類似度０以外の事例(489件)	45.8 %	15.75 %	1.204	0.499	0.068

注）「尺度Ａ」は上述の結果と同じである。まとめると、以下のようになる。

尺度	事例の分類	1 stの正解率	(1 st－last)の正解率	s_x/x	s_y	効果Ｅ
尺度Ａ	全事例（2003件）	77.2 %	8.89 %	0.125	0.419	0.149
	類似度０の事例(1514件)	87.4 %	6.61 %	0.010	0.332	0.074
	類似度０以外の事例(489件)	45.8 %	15.95 %	0.234	0.499	0.089

7．類似用例検索処理における翻訳正解率の向上効果

次に、上述の３つの尺度をＥＢＭＴシステムに組み込み、その翻訳正解率の向上効果を調べた。その実験のＥＢＭＴシステムにおける類似用例検索処理では、５件の類似用例を選択する処理が実行されるのであるが、同じ類似度を持つ用例の優先順位付けのために、これらの尺度が組み込まれた。

実験では、２５５０件の用例を持つ用例データベースから、１用例づつ抜き出してそれを入力とするジャックナイフ・テストを２５５０回実行した。実験結果は、選択用例５件内、及び、５件目を含む範囲内で同じ類似度を持つ用例がどのくらい発生するかに大きく依存する。表９に、同じ類似度を持つ用例の出現状況を示す。

この実験の結果（表１０参照）、全体の翻訳正解率は２％程度の向上に留まった。その原因は、５個すべてが類似度０である事例は、１３３２件中、１０８４件（事例全体の４２．５％）もあり、尺度による翻訳正解率の向上効果が低い類似度０の複数個の用例を持つ事例が多く存在したためである。

表９．類似用例５件内で、同じ類似度を持つ用例の出現状況

No	＼ 同じ類似度の用例数	５個	４個	３個	２個
1	５番目を含む同じ類似度を持つ用例の事例	1332	266	271	330
2	５番目を含まない同じ類似度を持つ用例の組数	－	135	238	472

注）「Ｎｏ２」は、１事例内に２個の同じ類似度の用例が２組存在する可能性があるからである。

表１０．類似用例５件を抽出する類似用例検索処理の結果

	翻訳正解率					
	全体	１番目	２番目	３番目	４番目	５番目
平均類似度	0.23	0.13	0.19	0.24	0.28	0.3
尺度なし	64.6 %	69.5 %	66.9 %	63.5 %	61.9 %	61.1 %
尺度Ａ	66.3 %	71.3 %	68.7 %	64.8 %	63.7 %	63.1 %
尺度Ｂ	66.4 %	71.2 %	68.9 %	64.8 %	63.9 %	63.3 %
尺度Ｃ	66.3 %	71.2 %	68.7 %	64.7 %	63.8 %	63.1 %

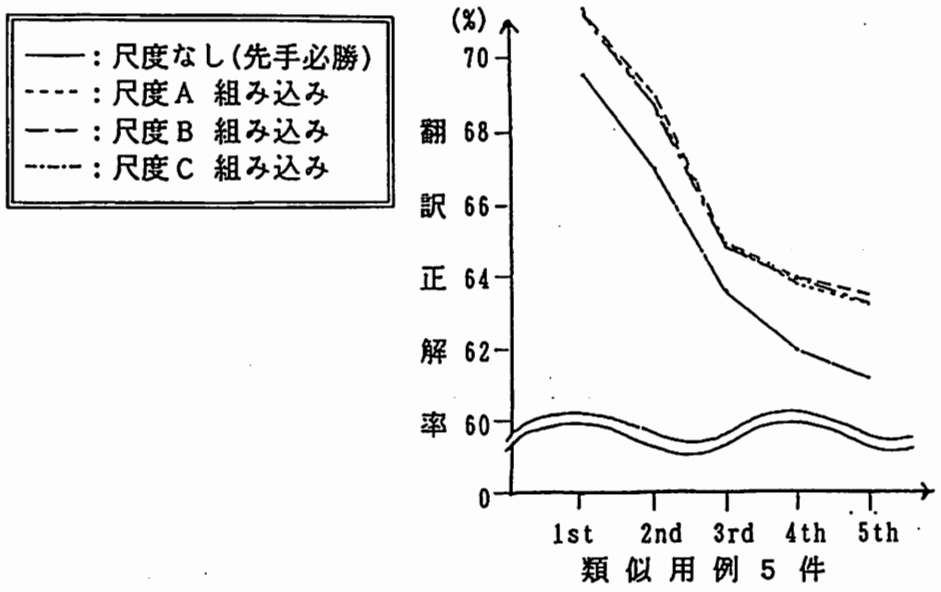


図８．類似用例に対する翻訳正解率の分布図
（２５５０件すべての事例を対象とする）

図８は、類似用例５件に対する翻訳正解率の分布を表したグラフである。これらの尺度が、常に、翻訳正解率を高める効果を持つことが分かる。実際の実施例を、［資料－２］に示す。

８．まとめ

本研究での成果をまとめる。まず最初に、同じ類似用例に対する類似選択におけるシャノンの情報量の特徴を明らかにした。それは、以下の２つの特徴である。

【特徴ａ】同じ類似度を持つ用例内で、各属性の情報量が入力の各属性の情報量と近接している用例ほど、翻訳正解率が高い。

【特徴ｂ】同じ類似度を持つ用例内で、用例に対する入力の一致成分（または、不一致成分）の情報量の総和が小さいほど、翻訳正解率が高い。

次に、これらの特徴に基づいた尺度が翻訳正解率の向上に効果的であることを示した。また、それらの尺度は、類似度が大きな（入力と用例との距離が大きい）事例ほど、翻訳正解率の向上効果も大きい傾向があることが分かった。

最後に、用例主導型機械翻訳（ＥＢＭＴ）システムの類似用例検索処理に、これらの尺度を組み込むことにより、翻訳正解率を高めることができることを確認した。

９．今後

今回の研究成果の利用目的として、以下の２点が考えられる。

(1) 一般の用例主導型機械翻訳（ＥＢＭＴ）システムの翻訳正解率を高めることができる。

(2) 類似用例検索処理を高速化する方法として部分検索手法を考えた場合、これらの尺度を使えば、用例の区分をさらに細分化できるため、検索対象（用例）の効果的な絞り込みが実現できる。

参考文献

[１] 隅田英一朗・飯田仁「用例主導型機械翻訳」，情報処理学会，自然言語処理研究会資料 82-5,1991,3,15

[２] 長尾真　著「言語工学」，昭晃堂，P32-P35

[３] 安本美典　他著「因子分析法」，培風館，P61-P67

[４] 大矢雅則　他著「確率論的エントロピー」，共立出版，P6-P10

[５] 電気学会　著「測定値の統計的処理」，電気学会出版，P62-P66

資料 — 1

以下で、「効果Ｅ」について説明する。

情報量の演算式を使用して、各事例において同じ類似度の用例を並び換え、そこでの翻訳正解率の分布を調べる。そして、演算式による用例の値をｘ、その時の翻訳正解率をｙとして、その分散（標準偏差）と標準化を伴う単回帰分析（最小二乗法による推定）による関数一次式を求め、翻訳成功率向上に寄与する効果を調べる。その関数一次式の傾きが、「効果Ｅ」である。

実際の実験では、各事例内の用例の集合を対象とはせず、演算式の値が最小のものを全事例から求めた用例集合を対象としたため、本来の目的での「効果Ｅ」を意味し得ない。なぜなら、比較対象の入力データが、すべての事象で異なっているからである。このため、演算式の値に対する翻訳成功率の偏りを表現している程度の解釈しかできない。

以下で、「効果Ｅ」の数式を求める。（文献［５］を参照）
関数一次式を以下のように決める。

$$y' = \beta_0 + \beta_1 x' \quad \therefore\ y' = \frac{y - \overline{y}}{s_y} \quad x' = \frac{x - \overline{x}}{s_x}$$

$\overline{x}$, $\overline{y}$：平均値　s_x, s_y：標準偏差

標準化を行うのは、与えられたデータを全て平均値０・標準偏差１に揃えるためである。これにより、β_0・β_1を直接比較できる。（文献［３］を参照）

［計算方法］

並び換えの尺度から求めた値ｘ、その時の正解率ｙ、データ数ｎとする。このｎ個のデータ（x_1, y_1），（x_2, y_2），・・・，（x_n, y_n）を母集団からの標本として扱う。

$$\text{平均}\overline{x}: \frac{\Sigma x_i}{n} \qquad \text{分散}s_x^2: \frac{\Sigma (x_i - \overline{x})^2}{n-1} \qquad x' = \frac{x - \overline{x}}{s_x}$$

$$\text{平均}\overline{y}: \frac{\Sigma y_i}{n} \qquad \text{分散}s_y^2: \frac{\Sigma (y_i - \overline{y})^2}{n-1} \qquad y' = \frac{y - \overline{y}}{s_y}$$

標準化のため、まず最初に（x_i, y_i）から（x_i', y_i'）に変換する。この時、

$$\overline{x}' = \overline{y}' = 0 \qquad s_x' = s_y' = 1 \qquad \Sigma x_i'^2 = \Sigma y_i'^2 = n - 1$$

が言える。

次に、正規方程式（注１）から、β_0・β_1の最小二乗推定値b_0・b_1を求めると、以下のようになる。

$$b_0 = \frac{\Sigma y_i'}{n} - \frac{\Sigma x_i'}{n} \cdot b_1 = \overline{y}' - \overline{x}' \cdot b_1 = 0$$

$$b_1 = \frac{\Sigma x_i' y_i' - \dfrac{(\Sigma x_i')(\Sigma y_i')}{n}}{\Sigma x_i'^2 - \dfrac{(\Sigma x_i')^2}{n}} = \frac{\Sigma x_i' y_i' - n\overline{x}\,\overline{y}}{\Sigma x_i'^2 - n\overline{x}^2} = \frac{\Sigma x_i' y_i'}{\Sigma x_i'^2}$$

よって、最小二乗法により、$y' = \dfrac{\Sigma x_i' y_i'}{\Sigma x_i'^2} x'$ が求まる。

ここで、成功率向上に寄与する効果Ｅを、以下のように定義する。

$$\text{効果E} := |b_1| = \left|\frac{\Sigma x_i' y_i'}{\Sigma x_i'^2}\right| = \left|\frac{\Sigma (x_i - \overline{x})(y_i - \overline{y})}{\sqrt{(\Sigma (x_i - \overline{x})^2)(\Sigma (y_i - \overline{y})^2)}}\right|$$

効果Ｅの値が大きいほど、成功率向上に寄与する効果が大きいと判断する。

注１）正規方程式とは、以下の連立一次方程式をいう。

$$\begin{cases} n b_0 + (\Sigma x_i') b_1 = \Sigma y_i' \\ (\Sigma x_i') b_0 + (\Sigma x_i'^2) b_1 = \Sigma x_i' y_i' \end{cases}$$

資料－２　類似用例５件を検索する『N1のN2』の用例主導型機械翻訳（ＥＢＭＴ）システムの類似用例検索処理に、尺度Ａを組み込んだ場合と通常の場合との比較事例を２件、以下に示す。

［テスト例－１］ 入力「会議の焦点」(kaigi no shouten) の場合

○ 尺度なし　：（3／5 "N2 of N1"）（2／5 "N1"）

```
DISTANCE TRPAT                  : JAPANESE ...[ENGLISH]
0.000000 (0 0 0 0 0 0 0)        : kono kaigi no shouten ...[the focus of the conference]
0.000000 (0 0 0 0 0 0 0)        : kaigi no shouten ...[the focus of the conference]
0.333333 (0 0 0 0 0 0 1/3 0)    : kono kaigi no mokuteki ...[the object of the conference]
0.333333 (0 0 0 0 0 0 1/3 0)    : kaigi no koto ...[your conference]
0.333333 (0 0 0 0 0 0 1/3 0)    : kaigi no hou ...[the conference]
```

○ 尺度Ａ 組み込み：（5／5 "N2 of N1"）

```
DISTANCE TRPAT                  :  Inf-A   : JAPANESE ...[ENGLISH]
0.000000 (0 0 0 0 0 0 0 0)      : 8.000000 :kono kaigi no shouten ...[the focus of the conference]
0.000000 (0 0 0 0 0 0 0 0)      : 8.000000 :kaigi no shouten ...[the focus of the conference]
0.333333 (0 0 0 0 0 0 1/3 0)    : 8.000000 :kono kaigi no mokuteki ...[the object of the conference]
0.333333 (0 0 0 0 0 0 1/3 0)    : 8.000000 :kaigi no mokuteki ...[the objective of conference]
0.333333 (0 0 0 0 0 0 1/3 0)    : 8.000000 :kaigi no mokuteki ...[purpose of conference]
```

［テスト例－２］ 入力「こちらの事情」(kochira no jijou) の場合

○ 尺度なし　：（4／5 "N1"）（1／5 "^N1 N2"）

```
DISTANCE TRPAT                  : JAPANESE ...[ENGLISH]
0.000000 (0 0 0 0 0 0 0)        : watakushi no jijou ...[my situation]
0.333333 (0 0 0 0 0 0 1/3 0)    : kochira no hou ...[We]
0.333333 (0 0 0 0 0 0 1/3 0)    : sochira no hou ...[you]
0.333333 (0 0 0 0 0 0 1/3 0)    : kochira no hou ...[we]
0.333333 (0 0 0 0 0 0 1/3 0)    : watakushi no hou ...[I]
```

○ 尺度Ａ 組み込み：（5／5 "^N1 N2"）

```
DISTANCE TRPAT                  :  Inf-A   : JAPANESE ...[ENGLISH]
0.000000 (0 0 0 0 0 0 0)        : 8.532331 :watakushi no jijou ...[my situation]
0.333333 (0 0 0 0 0 0 1/3 0)    : 8.011942 :sochira no touroku youshi ...[your registration form]
0.333333 (0 0 0 0 0 0 1/3 0)    : 8.533516 :watakushi no abusutorakuto ...[my abstract]
0.333333 (0 0 0 0 0 0 1/3 0)    : 8.533516 :watakushi no abusutorakuto ...[my abstract]
0.333333 (0 0 0 0 0 0 1/3 0)    : 8.533516 :watakushi no abusutorakuto ...[my abstract]
```