社団法人　電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
TECHNICAL REPORT OF IEICE.

SENSEVAL-2 日本語タスク

黒橋　禎夫†　　白井　清昭††

† 東京大学 大学院情報理工学系研究科
†† 北陸先端科学技術大学院大学 情報科学研究科

E-mail: †kuro@kc.t.u-tokyo.ac.jp, ††kshirai@jaist.ac.jp

あらまし　SENSEVAL-2 は語の意味的曖昧性解消のコンテストである．本稿では，SENSEVAL-2 の日本語タスクについて，タスクの概要，データ，コンテストの結果について述べる．日本語タスクでは，辞書タスクと翻訳タスクの 2 つのタスクを設定した．辞書タスクでは語の意味の区別を国語辞典によって定義し，翻訳タスクではこれを訳語選択，すなわち日本語単語に対する適切な英訳を選択する問題と定義した．両タスクとも評価テキストとして新聞記事を用いた．評価単語ののべ数は，辞書タスクが 10,000，翻訳タスクが 1,200 である．辞書タスクには 3 団体 7 システム，翻訳タスクには 5 団体 7 システムがそれぞれ参加した．

キーワード　語の意味的曖昧性解消，国語辞典，翻訳メモリ

SENSEVAL-2 Japanese Tasks

Sadao KUROHASHI† and Kiyoaki SHIRAI††

† Graduate School of Information Science and Technology, University of Tokyo
†† School of Information Science, Japan Advanced Institute of Science and Technology

E-mail: †kuro@kc.t.u-tokyo.ac.jp, ††kshirai@jaist.ac.jp

Abstract　SENSEVAL-2 is the second evaluation exercise for Word Sense Disambiguation programs. This paper describes two Japanese tasks in SENSEVAL-2: a dictionary task and a translation task. The dictionary task defined word senses according to a Japanese dictionary; the translation task defined word senses according to translation distinction. Test documents were newspaper articles in both tasks. The number of instances for evaluation was 10,000 in the dictionary task and 1,200 in the translation task. 7 systems of 3 organizations and 7 systems of 5 organizations participated in the dictionary/translation tasks respectively.

Key words　Word Sense Disambiguation, Japanese Dictionary, Translation Memory

1. はじめに

語の意味的曖昧性解消 (Word Sense Disambiguation; WSD) は機械翻訳，情報検索など，自然言語処理の多くの場面で必要となる基礎技術である [2]. SENSEVALは，ボランティアによるWSDのコンテストであり，WSDの共通の評価データを作成し，その上で様々なシステム・手法を比較することによってWSDの研究・技術を向上させることを目的としている．

第1回のSENSEVALは，1998年夏に英語，フランス語，イタリア語を対象として行われ，23研究グループが参加した [4]．SENSEVAL-2は，2001年春に日本語を含む9言語を対象に，37研究グループが参加して行われた(注1)．本稿ではSENSEVAL-2の中で行われた日本語タスクについて，タスクの概要，データ，コンテストの結果について述べる．

SENSEVAL-2では，タスクをlexical sample taskとall words taskに大別している．lexical sample taskは特定 (数十〜数百) の評価単語だけを対象とし，all word taskでは評価テキスト中のすべての単語を対象とする．日本語タスクでは，lexical sample taskとして，辞書タスクと翻訳タスクの2つのタスクを設定した．辞書タスクでは語の意味の区別 (曖昧性) を国語辞典によって定義し，翻訳タスクではこれを訳語選択によって定義した．なお，本稿では，評価単語の評価データ中での実際の出現を評価インスタンス，または単にインスタンスとよぶことにする．

SENSEVAL-2日本語タスクは次のようなスケジュールで実施した．

00/2/23	開催の呼びかけ
01/1/31	トライアルデータ公開
01/3/16	翻訳タスク Translation Memory 公開
01/5/11	評価データ公開
01/6/1	回答締切
01/7/6,7	ワークショップ (ACL01に併設)，結果発表

2. 辞書タスク

2.1 タスク概要

辞書タスクは，単語の語義を岩波国語辞典 [5] の語義立てによって定義し，WSDの正確さを競うタスクである．参加者は，テキスト中の評価インスタンスに対して，該当する語義を岩波国語辞典の語釈の中から選択し，その語釈に対応したID(以下，語義ID) を提出する．1つのインスタンスに対して複数の語義IDを返してもよい．また，インスタンスの意味がその語義IDである確率をつけて返してもよい．確率をつけずに複数の語義IDを回答した場合には，全ての語義IDの確率が等しいとして取り扱われる．

(注1)：http://www.sle.sharp.co.uk/senseval2/

むり【無理】
((名・ダナ)) 理を欠くこと。
㋐道理に反すること。「―が通れば道理が引っ込む」「君が怒るのは―もない（＝もっともだ）」。理由が立たないこと。「―な願い」
㋑行いにくいのに、押してすること。「―をして出掛ける」「仕事の―で病気になる」

図1　岩波国語辞典の「無理」の語釈文

テキストは毎日新聞の1994年の新聞記事を用いた．語義を決定する評価単語の数は100である．評価単語のそれぞれについて100語ずつ語義を決めるため，評価インスタンスの数は10,000である．

2.2 データ

本項では，辞書タスクで用いられた3つのデータ，岩波国語辞典，訓練データ，評価データについて述べる．

2.2.1 岩波国語辞典

岩波国語辞典の見出しの数は60,321，語義の総数は85,870であり，一見出し当たりの平均語義数は1.42である．岩波国語辞典の語釈の例を図1に示す．また，岩波国語辞典では，語義は階層構造を持つ(注2)．階層構造の最大の深さは3である．

辞書タスクでは，語義の定義として，形態素解析された岩波国語辞典の語釈文と，それに対応する語義IDが参加者に配布された．なお，語釈文の形態素解析結果は人手修正されている．

2.2.2 訓練データ

訓練データは，毎日新聞の1994年の3,000記事を解析したコーパスである．このコーパスに付与されている情報を以下にまとめる．

- 形態素情報 (分かち書き，品詞，読み，基本形)

コーパスに含まれる形態素数は880,000である．これらは人手修正されている．

- UDCコード

各記事には，テキストの分類カテゴリを表わす指標として，国際十進分類法 (Universal Decimal Classification, UDC) によるコード番号 [3] が付与されている．

(注2)：図1では，「理を欠くこと」という語義が㋐,㋑の語義の上位にある．

- 語義情報

各単語には，その単語の意味に該当する語義 ID が付与されている．但し，語義 ID はコーパスの全ての単語ではなく，1) 名詞，動詞，形容詞のいずれかである，2) 岩波国語辞典に見出しがある，3) 多義である，という条件を満たす単語のみに付与されている．その総数は 148,558 である．語義 ID は全て人手によって付与された．また，1 つの単語に語義 ID を付与した人は 1 人である．複数の人が同じ単語に語義 ID を付与し，それらを照合するといった作業は行われていない．

2.2.3 評価データ

評価データは，評価インスタンスとその正解となる語義 ID を含むテキストである．評価テキストとして毎日新聞の 1994 年の 2,130 記事を用いた．これらは訓練データの記事とは異なる．評価データに付与されている情報は以下の通りである．

- 形態素情報 (分かち書き，品詞)

これらは自動解析されたものである．訓練データとは異なり，人手による修正はされていない．

- UDC コード
- 語義情報 (正解データ)

評価インスタンスには正解となる語義 ID が付与されている．また，訓練データとは異なり，1 つのインスタンスに対して最低 2 人の人が語義 ID を付与している (詳細は 2.3 を参照)．もちろん，この情報はコンテストの際には参加者に配布されない．

2.3 正解データの作成

岩波国語辞典，訓練データ，評価データの付加情報のほとんどは，RWCP によって作成され，1997 年から既に公開されているデータである．訓練データの語義情報については [6]，それ以外の情報については [1] を参照していただきたい．これに対し，評価データの語義情報，すなわち正解となる語義 ID のデータは，今回のコンテストのために新たに作成した．ここでは，その作成過程ならびに概要について述べる．

2.3.1 評価単語の選定

評価単語を選定する際には，以下の点を考慮した．

- 評価単語の品詞は名詞または動詞とした．
- 訓練データにおける出現頻度が 50 以上の単語を評価単語とした．
- 訓練データにおける語義の頻度分布のエントロピー $E(w)$(式 (1)) を考慮した．

$$E(w) = -\sum_i p(s_i|w) \log p(s_i|w) \qquad (1)$$

表 1　評価単語数の内訳

	D_a	D_b	D_c	計
名詞	**10** (9.1/1.19)	**20** (3.7/0.723)	**20** (3.3/0.248)	**50** (4.6/0.627)
動詞	**10** (18/1.77)	**20** (6.7/0.728)	**20** (5.2/0.244)	**50** (8.3/0.743)
計	**20** (14/1.48)	**40** (5.2/0.725)	**40** (4.2/0.246)	**100** (6.5/0.685)

(平均語義数 / 平均エントロピー)

式 (1) において，$P(s_i|w)$ は単語 w の語義が s_i となる確率を表わす．$E(w)$ の値が大きい単語は，語義の頻度分布が一様であり，語義を決定することが比較的難しい単語であると考えられる．一方，$E(w)$ の値が小さい単語は，1 つの語義が集中して現われる傾向が強く，語義の決定も比較的易しいと考えられる．評価単語の選定の際には，$E(w)$ を WSD の難易度の目安とした．具体的には，高難易度の単語クラス D_a ($E(w) \geq 1$)，中難易度の単語クラス D_b ($0.5 \leq E(w) < 1$)，低難易度の単語クラス D_c ($E(w) < 0.5$) という 3 つの難易度クラスを設定し，それぞれのクラスから評価単語をまんべんなく選ぶようにした．

品詞別，難易度クラス別の評価単語の内訳を表 1 に示す．また，評価単語の岩波国語辞典における平均語義数，ならびに平均エントロピーも示した．

2.3.2 語義 ID の付与

10,000 語の評価インスタンスに対して，その単語の意味に該当する語義 ID を人手で付与した．語義 ID を付与した作業者は 6 名で，言語学や辞書編纂の知識をある程度持っている人達である．その手順を以下にまとめる．

（1） 2 人の人が独立に語義 ID を付与する．その際の大まかな指針は以下の通りである．

- 1 つの語義 ID を選択する．複数の語義 ID は選択しない．
- どの階層の語義 ID を選んでもよい．
- 岩波国語辞典の語釈の中に該当するものがなければ，UNASSIGNABLE(該当無し) とする．ただし，なるべく UNASSIGNABLE とすることは避け，岩波国語辞典の語釈の中から語義 ID を選択する．

（2） 2 者が選んだ語義 ID が一致していれば，それを正解の語義 ID とする．

（3） 2 者が選んだ語義 ID が一致していなければ，第 3 者がその中から正しいと思われるものを選択する．ただし，第 3 者が，2 者が選んだ語義 ID 以外の語義 ID が正しいと判断した場合には，3 人が選んだ

表 2　作業者の語義 ID の一致率

	D_a	D_b	D_c	(計)
名詞	0.809	0.786	0.957	0.859
動詞	0.699	0.896	0.922	0.867
(計)	0.754	0.841	0.939	0.863

3 つの語義 ID の全てを正解とする．

作業者 2 人の語義 ID が一致した割合を表 2 に示す．評価インスタンス全体における一致率は 86.3%である．名詞と動詞とで一致率を比較すると，それほど差が見られないことがわかる．また，名詞，動詞ともに，難易度の高いクラスの単語ほど一致率が低くなるが，その傾向は名詞よりも動詞の方が強いことがわかる．

語義 ID を選択する際，どの階層の語義 ID を選んでもよいとしたが，階層構造の末端以外の語義 ID が選択された単語の数は 94 であり，階層の上の語義 ID はあまり選ばれなかった．また，2 者の語義 ID が一致せず，第 3 者も違う語義 ID を選んだ単語の数は 28 であり，その全体に対する割合は 0.3%程度と非常に少なかった．

2.4 結　　果

辞書タスクには 3 団体 7 システムが参加した．参加団体とそのシステムの特徴は以下の通りである．いずれのシステムも訓練データを利用した教師あり学習を行っている．

- CRL1 ∼ CRL4 (通信総研)

Support Vector Machine(以下 SVM)，シンプルベイス，及びこれら 2 つの混合モデルである．

- Naist(奈良先端科学技術大学院大学)

学習アルゴリズムは SVM を用いている．また，PCA や ICA といった手法を用いて，有効な素性だけを残すように素性空間の圧縮を行っている．

- Titech1, Titech2 (東京工業大学)

学習アルゴリズムは決定リストである．素性として，評価インスタンスの文脈情報の他に，語釈文中の例文の情報も用いている．

SENSEVAL-2 では，全ての言語のタスクにおける共通の評価基準として，以下に述べる 3 つの評価基準がある (注3)．辞書タスクでも，この評価基準に従ってシステムの評価を行った．

- fine-grained scoring

正解の語義 ID とシステムの語義 ID が完全に一致していれば正解とする．

(注3)：評価基準の詳細については以下の URL を参照していただきたい．http://www.sle.sharp.co.uk/senseval2/Scoring/

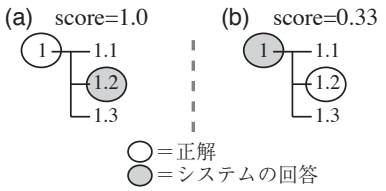

図 2　評価基準 (mixed-grained scoring)

- coarse-grained scoring

正解の語義 ID とシステムの語義 ID が，語義の階層構造の一番上の層で一致していれば正解とする．

- mixed-grained scoring

正解の語義 ID とシステムの語義 ID が完全に一致していなくても，語義の階層構造に従って部分的にスコアを与える方式で，fine-grained と coarse-grained の中間にあたる．語義の階層構造において，正解の語義 ID がシステムの語義 ID の親であるなら，正解とみなす (図 2 (a))．逆に，システムの語義 ID が正解の語義 ID の親であるなら，(1/正解の語義 ID の子の数) といった部分的なスコアを与える (図 2 (b))．

システムが複数の語義 ID を返したときには，各語義 ID の確率に従ってスコアの重み付き平均をとる．また，正解の語義 ID が複数ある場合は，正解の語義 ID 毎にスコアを計算し，その和を全体のスコアとする．

システムの結果を図 3 に示す．図 3 中の数値は，各システムの mixed-grained scoring によるスコアである．図 3 において，“Baseline” は訓練データにおける最頻出語義を選択したときのスコアを，“Agreement” は 2 人の作業者の語義 ID が一致した割合を示している．一番スコアが良かったのは CRL3 だが，どのシステムもベースラインを上回り，お互いのスコアの差も 3%程度で，それほど大きな差は見られなかった．

図 4 は，品詞別に見た各システムのスコア (mixed-grained) を示したグラフである．ベースラインを比べると，動詞の方が名詞よりも平均エントロピーが大きい (表 1) にも関わらず，約 3%ほどスコアが高い．これは，特にエントロピーの高い評価単語が動詞にいくつかあり，それらが動詞の平均エントロピーを大きくしているためと予想される．また，参加者のシステムを比べると，名詞よりも動詞の方がスコアの差が大きい．

図 5 は，難易度別に見た各システムのスコアを示したグラフである．クラス D_c の単語については，ベースラインや作業者の一致率も含めて，各システ

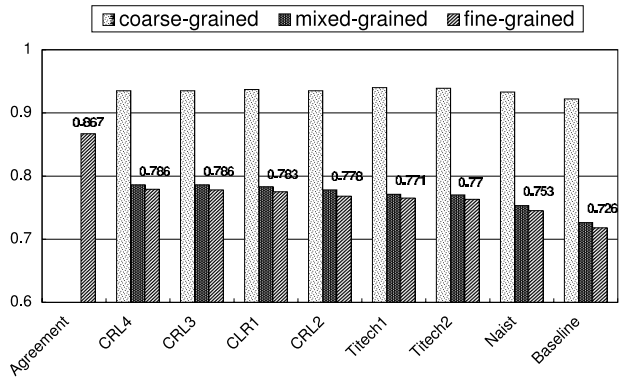

図 3　辞書タスクの結果

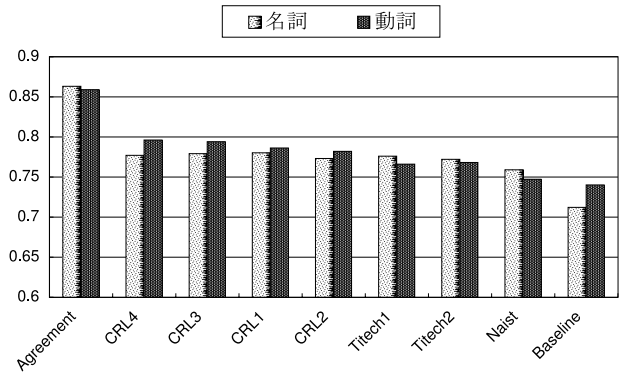

図 4　品詞別スコア (mixed-grained)

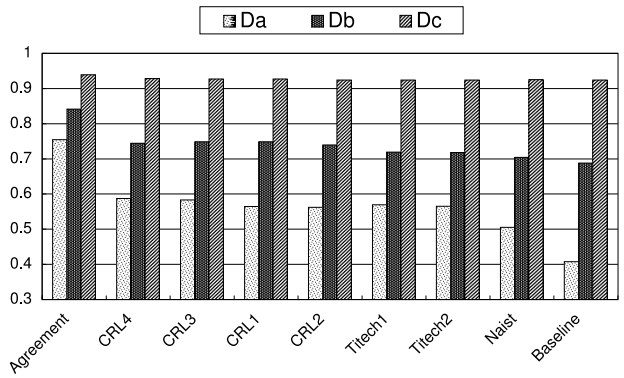

図 5　難易度別スコア (mixed-grained)

ムのスコアにほとんど差がない．これに対し，難易度の高い D_a や D_b の単語では顕著な差が見られる．

3. 翻訳タスク

3.1 タスク概要

意味的曖昧性をもつ語は，書き言葉の場合には，同綴り異義語 (homonym) と多義語 (polysemous word) に分けることができる．表音文字を用いる英語では多くの同綴り異義語があるのに対し，表意文字のある日本語では，漢字を用いる通常の書き言葉テキストに同綴り異義語はほとんどない(注4)．

(注4)：話し言葉であれば,「こうえん:講演，公園，公苑，公演，後援，好演，香煙...」のような同音異義語は多数存在する．

一般に同綴り異義語は意味の違いが明確であるのに対して，多義語の意味の区別は非常に微妙である．そのため，同綴り異義語が存在する英語では少なくともその部分で WSD の問題設定が明確であるのに対して，ほぼ多義語しか存在しない日本語では WSD の問題設定が非常に難しい．どのように意味を区別することが妥当か，有効であるかは，実際的にはアプリケーションが決まらなければ決まらない，ということがいえる．

そこで，日本語のもう一つのタスクとして，WSD を翻訳における訳語選択の問題と考える翻訳タスクを設定した．そこでは，対訳コーパスに基づく翻訳の考え方，すなわち，ある語についてさまざまな対訳用例を集めておき，新たな文の翻訳はその中から適切な用例を選択するという問題設定を行った．

そこで，まず，評価単語についてさまざまな対訳用例を集めた Translation Memory(TM) を作成し (図 6)，評価データ中の評価単語に対して TM の中からその翻訳に利用できる適切な対訳用例を選ぶことを課題とした．なお，既存の MT システムにも参加してもらい比較検討ができるように，評価テキストの翻訳を回答としてもよいこととした．

評価データは，辞書タスクと同じく，毎日新聞の 1994 年の新聞記事を用いた．40 語の評価単語を選択し (付録参照)，各評価単語について 30 個ずつの評価インスタンス (のべ 1,200 インスタンス) を用意した．

3.2 Translation Memory の作成

Translation Memory(TM) の作成は，

（1） 新聞の KWIC を参考に，対象語の典型的な日本語用例を列挙する (作業者 (言語学修士) の一次作業＋筆者らの確認)，

（2） 日本語用例を翻訳する (翻訳会社)

という 2 段階で行った．

KWIC には毎日新聞 9 年分を用い，形態素･文節解析を行い，対象語を含む文節の uni-gram, bi-gram(前後 2 種類)，tri-gram それぞれの頻度トップ 100 を取り出して作業者に提示した (図 7)．

作業者は，それらの中から対象語の典型的な日本語表現であると思われるものを抜き出し，それが文脈と独立に語の意味が明確であれば，そのまま用例とし (活用語の調整などは行う)，明確でなければ，新聞中での具体的使用例を参照し前後の表現を補って日本語用例を作成した．

次に，翻訳会社に依頼して日本語用例の翻訳を行った．翻訳が一意に決まらない場合には，訳が決まるように日本語用例を修正するか，それが難しい場合

無理 (むり)	
参加は無理だ	It is impossible to participate.
今から図書館の利用は無理だ	It is impossible to make use of the library in this hour.
今回の法案には無理がある	This bill is hard to pass.
彼が怒るのも無理はない	It is no wonder he got angry.
一番無理のない方法	the most natural way
無理を重ねる	to work too much
無理な話	unreasonable demand
無理な追い越し	passing by force
無理心中を図る	to commit a forced double suicide
...	...

図 6　Translation Memory の例

文節 uni-gram	文節 bi-gram		文節 tri-gram
597 無理な	151 無理は ない。	19 ことには 無理が	7 ことには 無理が ある。
551 無理が	138 無理が ある。	14 とても 無理。	6 求めるのは 無理が ある。
416 無理やり	106 無理も ない。	13 ことは 無理と	5 ことには 無理からぬ 理由が
413 無理に	101 無理 なく	10 求めるのは 無理が	5 嘆くのも 無理は ない。
403 無理を	67 無理の ない	10 とても 無理」と	5 同署は 無理心中と みている。
351 無理。	56 無理が ある」と	9 いうのは 無理が	4 しても 無理は ない。
...	...	...	...

図 7　TM 作成作業用の KWIC の例 (数字は表現の頻度)

には複数の翻訳を与え，メモ欄にその違いを説明するようにした．

最終的に 320 語に対して，合計 6920 用例 (1 語平均 21.6 用例) の TM を作成した (日本語用例の平均語数は 4.5，平均文節数は 2.6)．

3.3　評価データの作成と MT の評価

コンテストの評価データとして，TM の 320 語の中から評価単語 40 語を選び，各単語について評価インスタンスを 30 ずつ，合計 1200 インスタンスを選択した．

評価単語 40 語は，将来的に比較検討を行うことを考え，辞書タスクの評価単語 100 語の中から選択した．辞書タスクにおけるエントロピーに基づく難易度区分ごとに，名詞 D_a 5 語，D_b 10 語，D_c 5 語，動詞 D_a 5 語，D_b 10 語，D_c 5 語を選択した(注5)．

また，評価データの方も，辞書タスクとあわせ毎日新聞記事とした．辞書タスクでは各語 100 インスタンス，翻訳タスクでは各語 30 インスタンスなので，辞書タスクの 1 番目，4 番目，7 番目,... 90 番目の評価インスタンスを翻訳タスクでも評価インスタンスとした．

辞書タスクでは，正解データとして評価データの各インスタンスに語義 ID を 1 つ与えたのに対して，翻訳タスクでは，その語の翻訳に利用できる TM を 1 つまたは複数与えるという作業を行った．この作業は TM の翻訳と同じ翻訳会社に依頼して行った．正解は以下の基準で◎，○，△の 3 段階に分け，このような TM がない場合には「正解なし」とした．

◎　：翻訳に利用できる TM．日本語用例の品詞，時制，単複，微妙なニュアンス等は必ずしも一致する必要はない．

○　：評価単語のみに着目すれば妥当な訳語であるが，翻訳用例として使うことは望ましくない TM(例えば非常にまわりくどい表現になる)．

△　：評価単語のみに着目すれば妥当な訳語であるが，翻訳用例として直接は使えない TM．

1200 インスタンス中「正解なし」であったものは 34 インスタンス (0.03%)，1 インスタンスに対して◎，○，△は平均して 6.6，1.4，0.1 個，計 8.1 個与えられた．各インスタンスに対して TM 番号をランダムに選ぶとすれば，◎，○，△をすべて正解とする場合の正解率は 36.8%，◎のみを正解とする場合の正解率は 29.0% となった (次節の評価結果の Baseline)．

また，正解作成において，一部 (9 語 ×10 インスタンス) について 2 人の作業者が正解を与え，その一致率を計算した．この計算では，評価者 A(全体の評価担当者) の正解を基準とし，評価者 B の正解をランダムに 1 つ選んだ場合，それが評価者 A の正解と一致するかどうかを計算した．◎，○，△をすべて

(注5)：このような選択が可能となるように，辞書タスクと翻訳タスクの間で TM の単語選択，評価単語の選択を調整しながら行った．

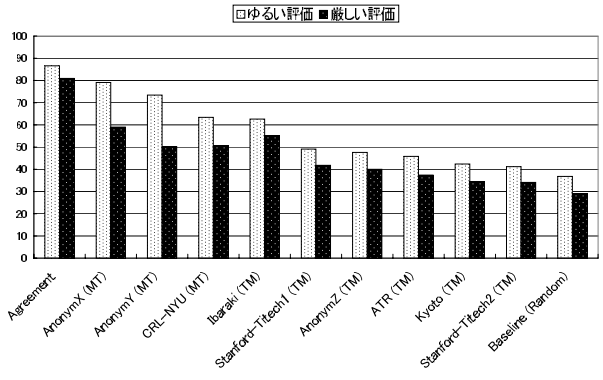

図 8　翻訳タスクの結果

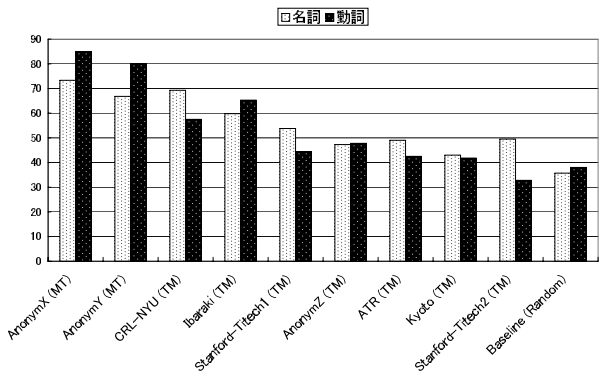

図 9　品詞別スコア

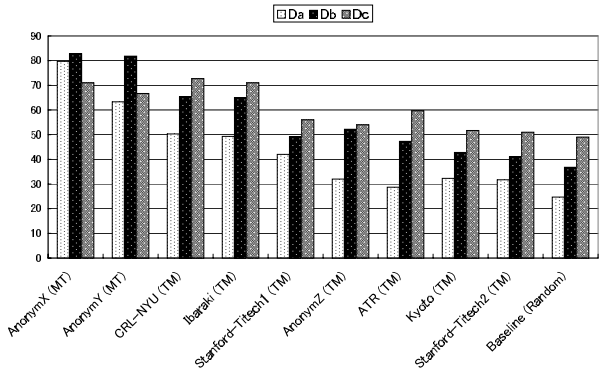

図 10　難易度別スコア

正解とした場合の一致率は 86.6%，◎だけを正解とした場合の一致率は 80.9%であった．

なお，翻訳タスクでは，評価データの翻訳を回答としてもよいとし，翻訳の単位は，評価インスタンスのみ，前後の句，文，文章全体のいずれでもよいとした．翻訳の回答が返されたものについては，コンテストの回答締切り後，やはり同様の翻訳会社に依頼して評価を◎，○，×の 3 段階で行った．ただし，文全体を翻訳している場合，全体の構文，まわりの表現とのバランスなどは考慮せず，評価単語の翻訳の妥当性のみを評価対象とした．

3.4 結　　果

翻訳タスクには 5 団体，7 システムが参加した．参加団体とそのシステムの特徴は以下の通りである．

- AnonymX，AnonymY (匿名)

それぞれ商用の機械翻訳システム．

- CRL-NYU (通信総研・ニューヨーク大学)

TM を英訳語ごとにクラスタリングし，そこに種々の類似用例を自動付与する．評価データと文字列類似度が非常に高い TM がある場合のみそれを回答とし，そうでなければ機械学習手法によって類似するクラスタの訳語を返す．

- Ibaraki (茨城大学)

評価単語について，毎日新聞コーパスから人手で学習データを作成し (1 評価単語あたり平均 170 インスタンス)，そこから前後 3 単語を見る決定リストを学習．

- Stanford-Titech1 (スタンフォード・東工大)

文字列類似度により評価データに最も一致する TM を返す．評価データ中の評価単語の他のインスタンスの情報も利用する．

- AnonymZ (匿名)

TM の日本語部分を形態素解析してから意味タグ列に変換し，これを学習データとして最大エントロピー法により学習．

- ATR

TM の日本語部分の形態素・構文解析を行い，評価単語の周辺の構文情報，意味情報をベクトル化し，評価データとのベクトルの cosine によって TM を評価．

- Kyoto (京都大学)

TM の日本語部分と評価インスタンスを含む文を，表現のバリエーションを吸収するボトムアップのマッチングシステムによって評価し，最も類似する TM を返す．

- Stanford-Titech2 (スタンフォード・東工大)

格解析を行い，格フレームとしての類似度で最も一致する TM を返す．

各システムの評価結果を図 8 に示す．左側は正解をゆるくとる (TM の◎，○，△，翻訳の◎，○をすべて正解とする) 場合の値，右側は正解を厳しくとる (◎のみ) 場合の値である．なお，TM には階層がないので，辞書タスクのような階層に基づく評価基準の区別 (fine,coarse,mixed) はない．図 9 に品詞別の結果，図 10 に難易度別の結果を示す．これらはいずれも正解をゆるくとった場合の値である．

グラフ中，Agreement，Baseline の値は前節で説明したとおりである．なお，システムは，適切な TM がないと判断した場合には「該当無し」(UNASSINABLE) と回答することができる．一方，評価データ

の方でも適切な TM がない場合には「正解なし」となっており，これらが一致した場合には正解，くい違った場合には不正解とした．

TM を選択するシステムについては，学習のためにデータ (情報) を増やしたシステムと，それ以外のシステムでスコアの開きがあった．MT システムと，TM 選択システムの比較は簡単にはできないが，MT システムにおけるノウハウの蓄積は十分評価される結果であった．一方，TM 選択システムの上位のスコアは MT システムのスコアと大差は無く，TM に基づく手法の今後の進展が期待できる．

最後に，TM の正解の与え方について参加者の一部に誤解があったことを説明しておきたい．参加者の一部は正解が 1 つまたは少数設定されると想定していが，運営側は利用可能な TM を (3 段階に区別するが) 広く解釈してすべて列挙した．運営側が「最適と判断される TM を 1 つ選択する」課題としたのは，実際に TM を用いる翻訳システムに近い状況を想定したためであり，正解が 1 つまたは少数であることを含意したものではなかった．参加者の一部からは，正解がそのように多数あるのであれば別の学習方法を考えたという意見があり，この点は運営上の反省点としたい．

4. おわりに

SENSEVAL-2 のようなコンテスト形式の評価型プロジェクトは国内外を問わずいくつか行われている．それらの成果のひとつとして，コンテストのために基礎的なデータが新たに作成され，コンテスト終了後に研究者らに公開されることが挙げられるだろう．日本語タスクでも，翻訳タスクの TM，翻訳・辞書タスクの正解データがコンテストのために新たに作成された．SENSEVAL-2 で使用されたデータは，日本語を含む全ての 9 言語について，SENSEVAL-2 の web サイトで公開されている (URL は 1 節の (注 1) を参照)．これらのデータが研究者らによって有効的に利用されることを願う．

謝辞

日本語タスクでは，評価テキストとして毎日新聞の新聞記事を利用させていただきました．新聞記事の利用に御協力いただきました毎日新聞社に感謝いたします．

日本語タスクの運営に数々の助言をいただいた東京工業大学の徳永健伸助教授，翻訳タスクの問題設定および TM 作成に協力いただいた通総研の内元清貴氏に深く感謝いたします．また，日本語タスクの種々のデータを作成して下さった作業者の皆様に感謝いたします．

文献

[1] Koiti Hasida et al. The RWC text databases. In *Proc. of the LREC*, pp. 457–462, 1998.
[2] Nancy Ide and Jean Veronis. Introduction to the special issue on word sense disambiguation. *Computational Linguistics*, Vol. 24, No. 1, pp. 1–40, 1998.
[3] 情報科学技術協会. 国際十進分類法 – 日本語中間版 – (第 3 版). 丸善, 1994.
[4] A. Kilgarriff and M. Palmer. Introduction to the special issue on senseval. *Computers and the Humanities*, Vol. 34, No. 1, pp. 1–13, 2000.
[5] 西尾実, 岩淵悦太郎, 水谷静夫. 岩波国語辞典 第五版. 岩波書店, 1994.
[6] 白井清昭ほか. 岩波国語辞典を利用した語義タグ付きテキストデータベースの作成. 情報処理学会自然言語処理研究会, Vol. 2001, No. 9, pp. 117–122, 2001.

付録

A. 評価単語

辞書タスクの評価単語の一覧を以下に示す．また，*のついた単語は翻訳タスクの評価単語でもある．

【名詞】
間, 頭, 一般*, 一方*, 今*, 意味*, 疑い, 男, 開発, 核*, 関係, 気持ち, 記録*, 技術, 現在, 交渉, 国内*, 言葉*, 子供, 午後, 市場, 市民*, 社会, 少年, 時間, 事業*, 時代*, 自分, 情報, 姿*, 精神, 対象, 代表, 近く*, 地方, 中心*, 手, 程度, 電話, 同日, 花*, 反対*, 場合*, 前*, 民間, 娘, 胸*, 目, もの, 問題*

【動詞】
与える*, 言う*, 受ける*, 訴える, 生まれる, 描く*, 思う, 買う*, かかる, 書く*, 変わる, 考える, 聞く*, 決まる, 決める, 来る, 加える, 超える*, 知る, 進む, 進める, 出す, 違う, 使う*, 作る*, 伝える*, 出来る, 出る*, 問う, 取る, 狙う, 残す, 乗る*, 入る, 図る*, 話す, 開く, 含む, 待つ*, まとめる, 守る*, 見せる*, 認める*, 見る, 迎える, 持つ*, 求める*, 読む, よる, 分かる

B. 評価テキストの例

評価テキストの例 (一部) を示す．評価インスタンスは ⟨head⟩ と ⟨/head⟩ で囲まれている．

... ある先生は、６０歳は人生のゴールではなく、これから第二の人生のスタート地点。４０代、５０代はそのための力の蓄えだとおっしゃった。　まだまだ老いてはいけない。せめて ⟨head⟩ 与え ⟨/head⟩ られた寿命までは生きて、二十一世紀をしっかり確かめてから旅立ちたい、と願うこのごろである。それにはやはり健康でいなければと心掛けている。...