多言語用例指南ツール：Kiwiの実験的評価

山本 真人[†]　**田中 久美子**[‡]　**中川 裕志**[‡]
[†]東京大学大学院 情報学環　[‡]東京大学 情報基盤センター
E-mail: {masato-y,kumiko,nakagawa}@r.dl.itc.u-tokyo.ac.jp

1　はじめに

本稿では我々が開発を行っている用例検索ツール，Kiwiの実験的評価を報告する．Kiwi は検索エンジンの結果を用いて多言語の語彙用例を調べるツールである [7]．語彙の用例を調べたい箇所に “*”（ワイルドカード）を入力すると，Kiwi はその語彙に関するデータを検索エンジンから得る．得たデータに統計処理を施しワイルドカードの位置に現れる用例候補を提示する．

Kiwi のアルゴリズムの中心部は，用例の切り出しとその順位付けである．Kiwi はまず検索エンジンの結果中からその語彙の周辺に現れる文字列を獲得し，ツリー構造化する．ここで文字列集合における分岐数という概念を導入する。分岐数とは、ある文字列の直後 (あるいは直前) に接続する文字の種類数として定義される。直感的にも分かるように分岐数は単語あるいは固定した言い回しの内部では先頭 (あるいは最後) 文字から進むにつれて減少し、単語や固定した言い回しが終わると、次には多数の単語がくる状態になるので増加する。この性質を利用し、Kiwi ではツリー中で分岐数が増加する直前までの文字列を候補として切り出し，さらに切り出した文字列を (頻度) $\times$ log(長さ $+1$) の評価関数を用いて順位付ける．これらの処理には，辞書や文法などの言語に依存するデータを一切用いていない．そのため，言語によらず用例を調べることができる．また言語データを検索エンジンから動的に得るため，多様な分野の最新用例を調べることができる．本稿では本ツールを用いた評価実験の結果を報告し，その有用性を検証する．

また，本稿では検索エンジンの結果が言語コーパスとして持つ意味を明らかにする．近年，多くの研究者が検索エンジンを用いた研究を行っている．このような研究においては，検索エンジンのアルゴリズムによりバイアスを受けた検索結果を，コーパスとして用いることの意味を正しく認識する必要がある．そこで異なる検索エンジンを用いた Kiwi の結果を比較することで，検索アルゴリズムの差異が結果に及ぼす影響を探る．

以下 2 章では熟語などの定型的用例，3 章では TREC Test を用いた評価実験により Kiwi の性能を評価する．4 章では Kiwi の検索ツールとしての総合的な有用性を，ユーザー実験を通じて明らかにする．

2　定型用例を用いた評価実験

まず本章では熟語や慣用表現などの定型用例を用い，Kiwi の性能を評価する．また AltaVista，Google，AllTheWeb と，異なる検索エンジンを用いた際の結果の比較も併せて行う．本実験により英語，フランス語，日本語を対象とした時の Kiwi の基本性能を検証する．

2.1　実験方法および評価尺度

本実験は以下の手順に従って進める．

1. 3 単語以上からなる熟語，慣用表現を辞書などから無作為に抽出
2. 抽出した語の先頭，中間，末尾のいずれか一部をワイルドカードで置換（以後，Kiwi に入力された語を質問，置き換えられた語を正解と呼ぶ）
3. 曖昧性のある質問を排除
4. 内容語を含まない質問を排除
5. 正解が一度も検索エンジンの結果中に現れない日本語の質問ーを排除
6. 質問を Kiwi に入力し検索
7. ベースラインとして，同様に AltaVista に直接質問を入力し検索
8. 評価尺度を計算

本実験では熟語集 [2] [3] や辞書 [6] からランダムで抽出した定型用例を用いる．質問は英語では各部位（先頭，中間，末尾）につき約 300 件，フランス語，日本語でそれぞれ約 100 件用意した．質問ーとした際に正解に曖昧性が生じる熟語は質問から除く．例えば，“* up with” のように “keep”，“come” と複数の正解をとる質問は排除した．また，“for a *” のように，内容語を含まない質問も除いた．

Google や AllTheWeb は分かち書きのない言語のインデクシングに形態素解析を用いており，単語やコロケーションが分割された検索語では正しい結果を返さない場合がある．この時 Kiwi は当然正解を提示することができない．しかし，これは Kiwi の性能による問題ではない．インデクシングの問題を含むクエリーを用いると，Kiwi の本来の用例処理性能を評価できない可能性がある．また，インデクシングを文字単位で行う AltaVista との比較を適切に行うため，インデクシングの問題を含む質問を排除することが望ましい．そこで，日本語の質問に関しては全ての検索エンジンの結果で，一度は正解が現れるもののみを用いることとした．

本実験では，検索エンジンから得るデータ量を 1000 マッチに設定した．マッチとは検索エンジンが提示するヒット数ではなく，検索結果中に質問と合致する文字列が現れる回数を表す．本稿ではこの回数をマッチ数と呼ぶ．検索エンジンから 1000 マッチ以下のデータしか得られない質問に関しては，得られたデータの範囲で処理を行う．評価尺度としては以下を用いた．

- N 位精度：正解が Kiwi の候補の N 位以内に提示される割合（N:1，10，1000）[単位：%]
- MRR：Mean Reciprocal Ranking

尚，本実験では，ベースラインとして AltaVista の結果上位から質問と合致する文字列の周辺に正解が現れて

表 1: 実験結果：英語　（N 位精度 単位：%）

	N:1	N:10	N:1000	MRR
Kiwi (末尾)	**A** 80.6	94.4	**B** 96.6	0.86
Kiwi (中間)	**A** 72.2	86.3	**B** 91.5	0.78
Kiwi (先頭)	**A** 78.3	95.9	**B** 97.5	0.85
baseline(末尾)	36.4	83.7	**C** 97.5	0.54
baseline(中間)	47.4	73.7	**C** 92.1	0.56
baseline(先頭)	34.4	79.5	**C** 99.2	0.49

表 2: 実験結果：フランス語　（N 位精度 単位：%）

	N:1	N:10	N:1000	MRR
Kiwi (末尾)	**A** 68.0	92.0	**B** 94.0	0.75
Kiwi (中間)	**A** 75.0	91.0	**B** 96.0	0.81
Kiwi (先頭)	**A** 63.0	87.0	**B** 93.0	0.72
baseline(末尾)	35.0	73.0	**C** 98.0	0.46
baseline(中間)	42.0	77.0	**C** 96.0	0.54
baseline(先頭)	39.0	71.0	**C** 98.0	0.48

表 3: 実験結果：日本語　（N 位精度 単位：%）

	N:1	N:10	N:1000	MRR
Kiwi (末尾)	**A** 70.1	94.1	**B** 100.0	0.80
Kiwi (中間)	**A** 86.2	98.6	**B** 99.3	0.92
Kiwi (先頭)	**A** 69.6	95.7	**B** 97.4	0.80
baseline(末尾)	41.0	76.1	**C** 100.0	0.51
baseline(中間)	69.0	94.5	**C** 100.0	0.77
baseline(先頭)	56.5	87.8	**C** 100.0	0.67

表 4: 検索エンジンごとの実験結果：英語

	N:1	N:10	N:1000	MRR
AllTheWeb(末尾)	74.6	94.4	98.1	0.82
Google (末尾)	76.5	96.2	98.1	0.83
AllTheWeb(中間)	70.7	88.5	92.6	0.77
Google (中間)	74.1	88.5	92.2	0.79
AllTheWeb(先頭)	75.8	94.3	97.5	0.82
Google (先頭)	77.0	94.7	97.1	0.84

いるかを併せて調べた．

2.2 実験結果

表 1，2，3 に検索エンジンに AltaVista を用いた際の実験結果を示す．表には Kiwi ，ベースラインそれぞれにおける N 位精度（N:1，10，1000）および MRR を記す．表によれば，1 位精度，10 位精度，および MRR では全ての言語において Kiwi がベースラインを上回った．特に英熟語の 1 位精度に関しては，検索エンジンでは 40%程度であるのに対し Kiwi では 80%程度と 2 倍近い．日本語ではベースラインでも比較的高い割合で正解を得ているが，この場合でも 1 位精度で 15%から 30%ほど高い精度を示している．Kiwi の 10 位精度は 90%から 95%であり，ベースラインと比較しても 5%から 20%ほど高い．結果から，Kiwi が行う集計能力の高さが分かる．

結果を比較すると英語が最も精度が高く，フランス語は比較的低い．これは検索エンジンから得られるデータ量の差異に原因があると考えられる．Global Reach [1] によれば，フランス語で記述されたページは Web 上の全ページの約 3.7%である．これは英語で記述されたページの約 10 分の 1，日本語の約 3 分の 1 である．本実験で検索エンジンから得られた平均マッチ数においても，英語で 600 〜 700 マッチ，日本語で 400 〜 500 マッチに対し，フランス語は 200 マッチ程であった．本実験の範囲では Kiwi の用例処理はデータが豊富であるほど良い結果を示すことが分かる．

2.3 エラー分析

Kiwi で正解が得られない，もしくは上位に提示されない要因を分析すると，以下の 3 つが考えられる．

一つ目は検索エンジンの結果にそもそも正解が含まれない場合である．これは Kiwi を原因とする問題ではない．これが要因となる割合は表中の $(100 - \mathbf{C})\%$ となる．つまり，表における **C** の値は Kiwi の上限を示す．

二つ目の要因はテストセットにある．本実験では，熟語集の用例を正解としたが，用例の観点での真の正解の定義は容易ではない．例えば本実験では “be anxious *” に対し “for” を正解としているが，Kiwi が第 1 位に提示した “to” を不正解とすることには疑問の余地がある．本来ならばこのようなクエリーは用いるべきではないが，全ての問題を排除することは難しい．そこで本研究においてはこの問題を含んだ上で実験を行っている．ここでテストセットが要因となる割合は表中の $(\mathbf{C} - \mathbf{B}) + (\mathbf{B} - \mathbf{A})\%$ に含まれる．

三つ目の要因は Kiwi の用例処理である．まず，切り出し処理が原因となる割合は (**C** - **B**)% である．Kiwi はツリー構造中の分岐情報を用いて用例を切り出すため [7]，用例の頻度が低い状況では切り出しに失敗する可能性がある．しかし表によればこの割合は英語，日本語の場合で 1 ，2%，フランス語でも 4，5%である．この割合にはテストセットの問題も含まれることを考えると，切り出し処理が要因となる場合は少ない．一方，順位付け処理が原因となる割合は (**B** - **A**)% であり，5%から 30%近い値となる．テストセットの問題とあいまっているとは言え，未だ改良の余地があると言える．

2.4 異なる検索エンジンの比較

本節では Google，AllTheWeb を用いて同様の実験を行った結果を比較する．表 4，5，6 に結果を示す．表によれば，英語においてはどの検索エンジンを用いてもほぼ同程度の精度が得られている．1 位精度では最大でも 5%，10 位精度では 2%程度の差しか見られない．フランス語では若干のばらつきが見られるが，精度の差は小さい．

しかし日本語では，AltaVista を用いた結果が突出している．これは検索エンジンのインデクシングアルゴリズムが原因である．すなわち Google, AlltheWeb では単語単位、AltaVista では文字単位のインデクシングであることによる．本実験では検索エンジンの結果中に正解が含まれない質問は排除した．しかしこの方法では，正解でインデクシングされていないページ内に偶然正解が現れるような質問は排除されない．この場合，正解が現れる頻度は本来の頻度と比較して非常に小さい．つまり Web 上での本来の用例の利用実態と異なるデータを

表 5: 検索エンジンごとの実験結果：フランス語

	N:1	N:10	N:1000	MRR
AllTheWeb(末尾)	57.0	87.0	91.0	0.69
Google (末尾)	55.0	82.0	85.0	0.65
AllTheWeb(中間)	70.0	92.0	97.0	0.78
Google (中間)	68.0	88.0	91.0	0.76
AllTheWeb(先頭)	61.0	88.0	96.0	0.69
Google (先頭)	58.0	85.0	90.0	0.67

表 6: 検索エンジンごとの実験結果：日本語

	N:1	N:10	N:1000	MRR
AllTheWeb(末尾)	60.7	83.8	94.0	0.72
Google (末尾)	60.7	84.6	91.4	0.70
AllTheWeb(中間)	70.3	89.7	99.3	0.78
Google (中間)	78.6	93.1	98.6	0.84
AllTheWeb(先頭)	60.0	85.2	90.4	0.70
Google (先頭)	57.4	87.0	89.6	0.69

Kiwi は用いることとなる．この問題が原因となり，文字単位でインデクシングを行わない二つの検索エンジンの結果で精度が低下していた．このことから分かち書きのない言語で検索エンジンの結果を用いる際は，ランキングアルゴリズムよりもむしろインデクシングアルゴリズムに留意する必要があると言える．

3 TREC を用いた評価実験

本章では最新用例や専門用語を対象とした時の Kiwi の性能を検証する．本実験では，TREC [4]（Text REtrieval Conference） 2002 の Question and Answering track からランダムで抽出した 50 題を用いた．

3.1 実験方法および評価尺度

本実験は以下の手順に従って進める．

1. TREC QA track よりランダムに問題を抽出
2. 問題文を疑問文から平叙文へと変換したものを質問として Kiwi で検索
3. 正解が得られなかった場合，問題文中の単語を組み合わせて生成した質問を用いて検索
4. 評価尺度を計算

以下に問題例を示す．

- When did the shootings at Columbine happen?
- What is the scientific name for tobacco?
- What river is called China's Sorrow?

本実験で用いた問題はいずれも近年の出来事や，専門的な内容が正解となっている．また質問応答的な問題であるため，より現実に即した場面における有用性の検証となる．

TREC に記載された問題は疑問文である．そこで，疑問文を平叙文の形式に書き換え，正解が現れる位置にワイルドカードを記述したものを質問とした．正解が得られなかった場合，TREC の問題文中の単語を組み合わせ質問を再生成する．この場合，複数の質問が生成される可能性がある．本実験においては，考えうる質問を全て作成し，最も成績の良いものを用いる．

表 7: 実験結果：TREC

平叙文 （%）	組み合わせ （%）	平均順位	MRR
52.0	74.0	1.85	0.77

表 8: 実験結果：ユーザー実験

	時間 (分)		クリック数		自信度	
	$\bar{x}$	S	$\bar{x}$	S	$\bar{x}$	S
Kiwi	1.01	0.77	3.40	2.86	4.64	0.67
検索エンジン	1.40	1.18	7.04	6.24	4.06	1.11

本実験では評価尺度として，Kiwi が正解を提示した問題の割合，提示した場合に正解が現れる順位の平均，そして MRR を用いた．

3.2 実験結果

表 7 に実験結果を示す．結果によれば，問題文を単に平叙文に変換することで半数以上で正解を得た．単語の組み合わせによって生成した質問では 74%の割合で正解を提示した．また正解が得られた場合では，結果の上位に正解が提示されている．結果から，TREC のように正解が現代的，専門的であり，難易度が高い問題に対しても，Kiwi が有用であることが分かる．

4 ユーザー実験

本章では，ツールとしての Kiwi の総合的な有用性を，ユーザー実験により検証する．被験者は，質問応答的な問題計 32 問に対し検索エンジン，もしくは Kiwi を用いて解答する．各問題につき解答時間，クリック数，解答に対する自信度を計測し，結果の比較を行った．

4.1 実験方法および評価尺度

本実験は以下の手順に従って進める．

1. J.M.Spool ら [5] の問題設計方針に基づき，問題を作成
2. 問題を I 群，II 群に分割
3. 被験者を A グループ，B グループに分割
4. A グループの被験者は I 群に検索エンジン，II 群に Kiwi を用いて解答
5. B グループの被験者は I 群に Kiwi ，II 群に検索エンジンを用いて解答
6. 評価尺度を計測

本実験では，Spool ら [5] が行う Web Site Usability Testing の問題設計方針に基づいて作成した問題，計 32 問を用いた．以下に問題例を示す．

- 1 ヤードは何メートル？
- オリオン座は冬の星座？夏の星座？
- 西武ライオンズの松坂投手の決め球は？

被験者には，インターネットを頻繁に利用している理工系の学生 20 名を用いた．Kiwi の利用経験がある者は

図 1: Kiwi と検索エンジンの結果比較：解答時間

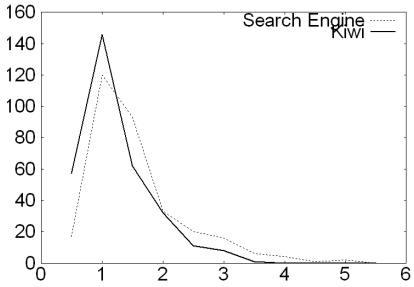
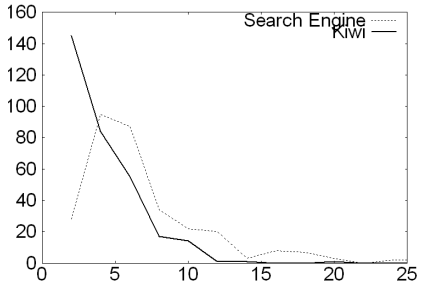


図 2: Kiwi と検索エンジンの結果比較：クリック数



内 4 名であった．被験者のグループは，Kiwi の経験者をまず均等に分け，他の被験者はランダムでグループ分けを行った．被験者は，それぞれ検索エンジンもしくは Kiwi を用いて問題に解答し，解答までに要した時間，クリック数，およびその解答に対する自信度を 5 段階評価で記入させ，結果を比較した．

4.2 実験結果

表 8 に，それぞれの尺度に関して，平均値（$\bar{x}$），標準偏差（S）を示した．表によると，Kiwi を用いた方がより短時間，少ないクリック数で，自信度の高い解答を得ていることが分かる．3 つの尺度を有意水準 1%で Welch の検定を行ったところ，全ての尺度で有意差が認められた．注目すべき点として，Kiwi を用いた場合では結果の標準偏差が小さいことが挙げられる．つまり，利用者によらず解答に辿り着く時間、クリック数の差異が小さい．これは Kiwi は個人の検索技能の影響が小さいことを意味する．この点は検索ツールとして重要な特性であると言える．

図 1，2，3 は，それぞれの尺度における Kiwi と検索エンジンの結果の分布を表したものである．横軸はそれぞれ，図 1 は解答時間（分），図 2 はクリック数（回），図 3 は回答への自信度（5 段階評価）を表し，縦軸はサンプル数を表す．図より，特に Kiwi を用いた場合少ないクリック数で解答を得られていることが分かる．検索エンジンを用いる場合，複数の候補を得るためには多くのページを調べる必要がある．これに対し，統計処理を行い候補の一覧を提示する Kiwi では，一度の検索で多くの候補が得られるためであると考えられる．

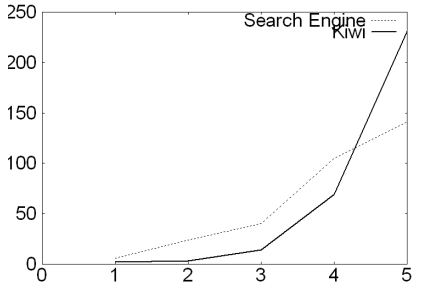
一方，解答時間の尺度では比較的差が小さい．Kiwi は検索エンジンの大量の結果を動的に得ており，そのダウンロードに多くの時間を必要とするためと考えられる．今後の課題としては，検索エンジンと同一サーバー上で Kiwi を実行し，ダウンロード時間を軽減することが考えられる．

自信度の尺度では Kiwi がより良い結果を示した．実験後のアンケートによれば，多くの被験者が Kiwi の集計能力を高く評価しており，自信度の結果として表れていたと言える．

5 おわりに

本論文では，Web 上の検索エンジンの結果を用いて語彙の用例を調べるツール，Kiwi の実験的評価を報告した．熟語や慣用表現などの定型用例を対象とする場合で

図 3: Kiwi と検索エンジンの結果比較：解答への自信度



は，85%から 98%の割合で 10 位以内に正解が提示された．対象とした言語ごとの精度を比較したところ，Web 上における文書の量が多い英語で最も良い成績を示していた．次に TREC の問題を用いた実験を行った結果，現代的な用例や専門用語を得ることができた．これらの実験から，Kiwi の統計処理能力の高さと潜在的な有用性が示された．また，ユーザー実験の結果からも Kiwi が検索ツールとして高い性能を示すことが得られた．一般に Web 上のデータはノイズが多いとされている．しかしながら，検索エンジンの結果に統計処理を施すことにより，用例検索のコーパスとして有益となりうることが実験より明らかとなった．今後は，処理の高速化を中心にシステムの最適化を行う．また，Question and Answering や文章校正など，より高度なタスクへ本ツールを応用していくことを試みる．

参考文献

[1] GlobalStats. Global reach community, 2002.
[2] K. Hisamatsu. 仏検 2 級・3 級対応 フランス語重要表現・熟語集. 駿河台出版社, 2001.
[3] T. Kanbe. TOEFL 英熟語 850. 旺文社, 2001.
[4] NIST, IAD, and ARDA. Text retrieval conference (trec) home page, 2003.
[5] J. M. Spool and J Pool. *Web Site Usability*. Morgan Kaufmann Publishing, 1998.
[6] 三省堂編修所. 三省堂実用ことわざの辞典. 三省堂, 2002.
[7] 山本真人, 田中久美子, 中川裕志. 検索エンジンに基づく多言語用例指南ツール：kiwi. 言語処理学会大会論文集, pp. 654–657, 2003.