Microsoft IME のユーザーフィードバックに基づく品質改善

大附克年　　**石橋紀子**
マイクロソフト ディベロップメント株式会社
東京都調布市調布ヶ丘1-18-1
katsutoshi.ohtsuki@microsoft.com
norikois@microsoft.com

鈴木久美
Microsoft Research
One Microsoft Way
Redmond WA 98052 USA
hisamis@microsoft.com

概要

日本語入力ソフトにおける変換精度は、ユーザーの生産性に直結する重要な指標の一つであるが、日本語入力ソフトを使って入力されるテキストは多岐にわたり、どれだけ大規模なテストデータを用意しても、実際にユーザーが入力するテキストをすべてカバーすることはできない。Microsoft IME では、ユーザーの同意に基づいて誤変換データや単語登録データの収集を行っており、現在、一日当たり約 10 万件の誤変換データと、約 2,000 件の単語登録データがユーザーから送信されてきている。本稿では、これらのユーザーフィードバックの仕組みによって、実際にどのようなデータが収集されているのか、また、収集されたデータをどのように品質改善に活用しているかを紹介する。

1　はじめに

マイクロソフトの日本語入力ソフト Microsoft IME は日本で一番使われている入力システムであり、そのユーザーベースは非常に多岐にわたりすそ野が広い。それゆえ、誤変換にもさまざまなものがあり、それらを収集・分析することは、開発の実験室で把握しきれないユーザーのニーズを知ることになり、製品開発をすすめていく上で非常に重要である。マイクロソフトではカスタマーエクスペリエンス向上プログラム (CEIP, Customer Experience Improvement Program) として、様々な場面でのユーザーからのフィードバックを集めている[1]。IME に関しては、CEIP を通じて、たとえばあるプロセスで変換誤りが何%あったかなどの統計データを取得している。また、IME に関する文字列を含む詳細なフィードバックを、IME Watson と呼ばれるプロセスを通じてあわせて収集している。IME Watson の中心となるフィードバックは、IME 誤変換レポート、単語登録レポートおよび学習データの三種類からなっている。本稿でユーザーフィードバックとして対象としているのは、CEIP および IME Watson に集まったデータである。

2　IME Watson レポートの仕組み

誤変換レポートおよび単語登録レポートの収集は、Microsoft Office IME 2003 (IME 2003) から、学習データの収集は、Microsoft Office IME 2007 (IME 2007) から行われており、最新の Microsoft Office IME 2010 (IME 2010) でも続けられている。誤変換レポートは、誤変換データをユーザーの環境に蓄積しておき、一定量たまるとユーザーの明示的な同意に基づいて、マイクロソフトに送付する仕組みになっている。誤変換レポート送付の画面を図 1 に示す。この画面は、誤変換データが一定量たまったとき、あるいは言語バーの「ツール」から「誤変換レポートを送信」を選択することによってアクセスできる。ユーザーは送付するデータをこの画面でチェックし、削除・編集することも可能である。

学習データはさらに、「誤変換レポート」の「設定」で、「学習データを送信する」を選択したときにのみ送信される。学習データの送信設定画面を図 2 に示す。単語登録レポートは、言語バーの「ツール」「単語の登録」で、「登録と同時に単語情報を送信する」をオンにしたときのみ、送信される仕組みになっている。

このような仕組みを通じて、実際にどのようなフィードバックデータが集まるかは、変換結果に依存するため、当該の IME のバージョンによって異なる。本稿では最新の Microsoft IME である、IME 2010 [2] に関するフィードバックデータの分析を主眼とするが、解析の手法やその使用法はそれ以前に収集されたデータにもほぼ同じように当てはまる。なお、Microsoft IME は、IME 2007 から統計的言語モデルを採用しており、Section 4 でのデータの活用法は、統計的言語モデルの訓練データに関する話題を含む。また、IME 2010 では、辞書アップデートによって後述のプロセスを通じて改善された辞書および言語モデルを自動配布するようになっている。

3　フィードバックデータの分析

3.1　CEIP 経由で収集されるデータ

CEIP では、文字列情報は集めていないため、IME で候補を開いた回数や Backspace キーを使用した回数な

[1] http://www.microsoft.com/products/ceip/ja-ja/default.mspx.

[2] http://www.microsoft.com/japan/office/2010/ime/

どの統計情報のみが送信される。CEIP 経由のデータの中で IME にとって一番有益な情報は、文字誤り率 (Character Error Rate, CER) であるが、このほかにも、CEIP 経由のデータから、たとえば「平均してキーストローク 10 回に対して 1 回変換されている」「平均すると 2 文字打つたびに 1 回 Backspace キーが使用されている」などの IME に関して有益な情報が抽出できている。Backspace キーの使用については、その原因について別途調査を行った結果、タイプミス、推敲による書き直し、変換誤りがそれぞれ三分の一を占めていた。

3.2 IME Watson レポート

IME Watson レポートには現在、一日当たり約 10 万件の誤変換データと約 2,000 件の単語登録データが送信されてきている。時期的な変動としては、週末やお盆、年末年始の時期に少なくなる傾向が観察され、ビジネスユーザーからのレポートが多いと推測される。

3.2.1 単語登録レポート

単語登録レポートには、単語登録ダイアログからユーザーが登録した単語の表記、よみ、品詞、ユーザーコメントなどが含まれる。登録単語は、おおまかに下記のように分類することができる。

- 分野特有の単語:医療用語、建築用語など
  例) アスベスト粉塵濃度測定(あすべすとふんじんのうどそくてい)、結晶化学(けっしょうかがく)、アミノ酸残基(あみのさんざんき)
- 顔文字
- 人名
  例) 真之(さねゆき)、髙橋(たかはし)、春風亭昇太(しゅんぷうていしょうた)
- 短縮よみ
  例) 株式会社(かぶ)
- その他固有名詞
  例) 元氣プロジェクト(げんきぷろじぇくと)、いかるが牛乳(いかるがぎゅうにゅう)

登録単語の中で最も多いのが、分野特有の単語で全体の 4 割強を占める。顔文字と人名がそれに次いで多く、それぞれ約 2 割となっており、これらの上位三分類だけで登録単語全体の約 8 割を占めている。

3.2.2 誤変換データ

誤変換データは、ユーザーが変換キーを押して得られた変換結果と異なる文字列を確定したときに収集される。具体的には、候補リストを開いてそこから単語を選んだ場合や文節区切りを変更した場合などである。収集は文節単位で行っており、誤変換とその前後一文節が収集の対象となる。文節区切りを変えた場合には、その部分と前後一文節が収集される。たとえば、最初の候補の文節区切りが

```
A  B  Cc  Dd  E  F
```

で、CcとDdの文節境界を変更して

```
A  B  C  cDd  E  F
```

のようにした場合、文節区切りを変えた部分 (CcDd) ＋前後 1 文節の、B～E までが収集対象となる。

表 1 に誤変換のタイプごとの代表的な例と、約 19 か月分 (2007年1月～2008年7月) のデータにおける相対頻度を示す。

表1: 誤変換のタイプと代表的な例

タイプ	結果	正解	頻度
単語の誤変換	小説の県	小説の件	55%
未登録語	黒木名さ	黒木メイサ	9%
文字列の誤変換	取って	撮って	8%
文節区切りミス	の浮	納期	6%
カタカナ未登録語	顧問図	コモンズ	4%
その他			18%

この表からわかるように、過半数の誤変換が、辞書に登録されている語が出てきてほしいコンテキストで出てこないという、語の曖昧性に由来する誤変換である。ただし、このような曖昧性に由来する誤変換には、表記ゆれ(たとえば「友達」と「友だち」などのペア)も含まれており、IME の誤変換の自動解析を難しくしている。

3.2.3 学習データ

学習データとは、確定した単語の統計量を指し、第一候補が正解だったものもそうでなかったものも含めて、単語 ID とその頻度情報からなっている。

IME 2010 の出荷時 (2010年5月) から 6 か月分の学習データ(延べ約 8,800 万語)を調べたところ、一番多いのは当然ながら助詞の類で、「の」「、」「に」「を」「が」「は」の順である。内容語では、「お願い」「確認」「必要」「情報」などの語が高位にランクしており、ここでもビジネスユーザーからのレポートが多いことをうかがわせる。

4 フィードバックデータの活用

この節では、個々のデータ種別の活用法とその評価について述べる。

4.1 データ種別の活用法

4.1.1 CEIP からのデータ

CEIP を通じて、バージョンの違いやアップデートによる CER の推移を知ることができる。母集団が同じではないので厳密な比較とは言えないが、集まっているデータの量が多ければ傾向を推し量ることができる。例えば、IME 2010 の CER は IME 2007 の半分以下であるということを CEIP のデータが示している。

一方で、実験室での評価結果と CEIP のデータに乖離がある場合には、実験室の評価方法が実際のユーザーの入力を反映していなかったり、何らかの理由で CEIPのデータが適切に収集されていなかったりといったことが考えられ、誤変換データや学習データなどを手がかりに原因を調査することになる。

4.1.2 登録単語

ユーザーから送られてくる単語の情報を集計し、累積の件数が一定数を超えたものは、単語登録しないユーザーも使う可能性のある単語と考えて、辞書への追加登録を行う。その際、短縮よみと顔文字は除外している。また、誤った単語や差別的な表現などについても除外するよう人手によるチェックを行っている。こうして追加登録が行われた辞書は辞書アップデートによってユーザーに提供される。

4.1.3 誤変換レポート

誤変換データについても、登録単語と同様に、累積の件数の多いものから優先的に修正して、辞書アップデートを通じて新しい辞書と言語モデルを配布している。誤変換を修正するためには、まず誤変換の原因を切り分け、その原因ごとに個々の修正作業を行っていく。辞書エントリに起因する誤変換に対してはエントリの追加や属性の変更を行い、言語モデルが原因の場合には、訓練コーパスの追加や誤りの修正を行う。ここでも、打ち間違いや誤った読みから変換しているものは除外の対象としている。

4.1.4 学習データ

学習データからは、実環境で使われている単語の頻度分布を知ることができる。この頻度分布を、言語モデルの訓練コーパスにおける単語の頻度分布と比較することにより、訓練コーパスで相対的に不足している単語を抽出することができる。例えば、同音異表記語に着目して、ある同音語のグループ H における異表記の出現分布が訓練コーパスと学習データとの間でどれだけ歪んでいるかを下記の $D(H)$ のように定義することができる。

$$D(H) = \sum_{h_k \in H} \left| P_{corpus}(h_k) - P_{user}(h_k) \right|$$

$$P_{corpus}(h_k) = \frac{C_{corpus}(h_k)}{\sum_{h_i \in H} C_{corpus}(h_i)}$$

$$P_{user}(h_k) = \frac{C_{user}(h_k)}{\sum_{h_i \in H} C_{user}(h_i)}$$

IME 2010 では、この歪みの値が大きい同音語のグループの中で、訓練コーパスにおける出現頻度が学習データに対して相対的に少ない単語に着目し、それらの単語を含む文を収集して、新たに訓練コーパスに追加した。この訓練セットの追加により、同音異表記語を含む評価セットに対する変換精度が改善することを確認している。

4.2 フィードバックに基づく改善とその評価

上述のように、辞書に新たな単語を追加したり、ある特定の単語グループに限って訓練コーパスを操作したりした場合、懸念されるのはその評価コーパス全般への影響である。つまり、ある特定の分野の変換精度向上が、別の分野で副作用を生んでしまうような状況が考えられる。したがって、辞書や訓練コーパスの更新は、常に評価セット全体に対する性能評価を伴っていなくてはならない。IME 2010 の辞書アップデートでは、辞書の配布に先立って、さまざまなタイプの文を集めた評価コーパスに対して数十項目に及ぶ副作用チェックを行っている。

出荷時（2010年5月）の IME 2010 (May2010) と2010年12月にリリースした辞書アップデートを適用した IME 2010 (Dec2010) を複数の評価コーパスに対して評価した結果を表2に示す。IME 2010 では、出荷時以降、2010年12月を含む複数回の辞書アップデートによって、ユーザーフィードバックおよびその他のソースからの情報に基づいて約 12,000 語を追加している。評価セットは、新しい単語を含むものとして、辞書アップデートで追加された単語を含む 1,050 文 (Added) と2010年7月から12月の間に検索エンジン Bing の「気になる言葉」[3]で紹介された単語 353 語 (Bing) を用いた。副作用をチェックするための一般的な評価コーパスとしては、さまざまな内容と文体を含む約 12 万文 (General) を用いた。

表 2 からわかるように、新しい単語を含む評価コーパス(Added, Bing) に対しては辞書アップデートにより大幅にCERが改善している。一方で、一般的なコーパス (General) に対しては CER はほとんど変わっておらず、単語の追加による大きな性能劣化は起きていないと考えられる。

表1: IME 2010 辞書アップデートの評価

評価データ	CER		CER
	May2010	Dec2010	削減率
Added	12.3%	5.5%	55.3%
Bing	21.8%	12.5%	42.6%
General	3.82%	3.79%	0.008%

5 おわりに

本稿では、Microsoft IME が継続的に収集しているユーザーフィードバックデータについて紹介し、その分析と活用法を概観した。今後もユーザーフィードバックデータを積極的に製品開発に取り入れていくが、一方でより多くのユーザーからフィードバックを集められるような働きかけもしていきたい。また、フィードバックデータの分析や IME のアップデートに関する情報は、随時 IME チームのブログ[4]でも紹介している。

[3] http://keyword.jp.msn.com/kinikoto/

[4] http://blogs.technet.com/b/ime/

図1: 誤変換レポートの送信画面

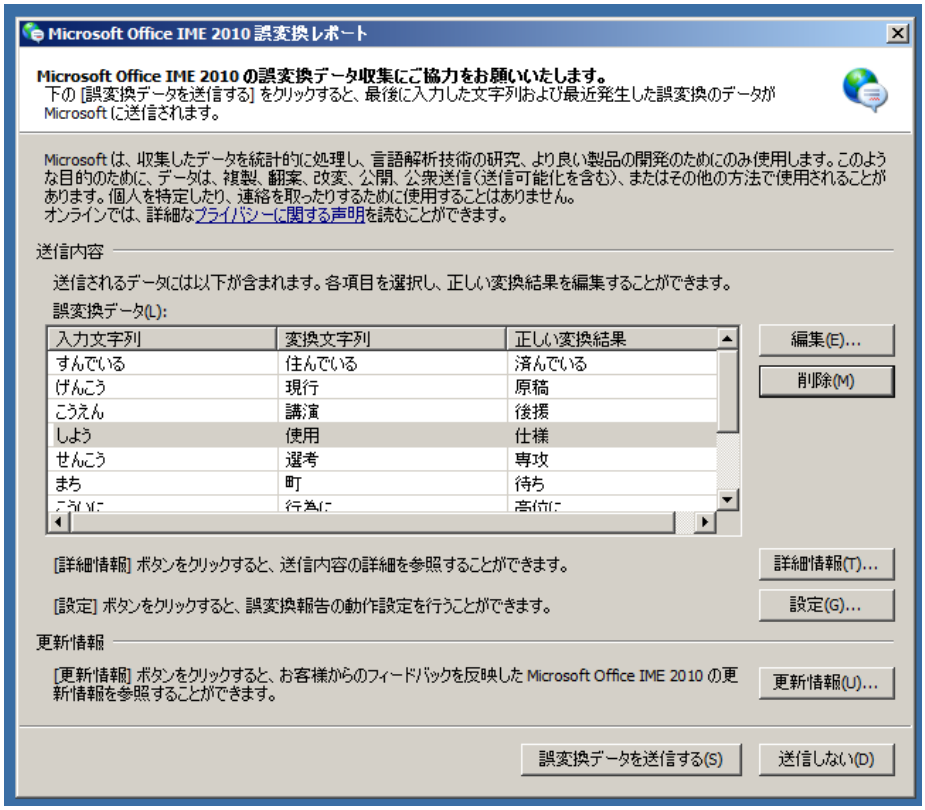

図2: 学習データの送信設定画面