文の構造を明示的に指定・表示することによる異言語間コミュニケーション

Translingual Communication by Explicit Specification and Presentation of Sentence Structures

五十嵐 健夫*

概要. 機械翻訳技術は、統計的な手法の発展などにより実用レベルに近づいているが、長い文の内部における係り受け構造の推測などは依然として難しい問題として残されている。本稿では、このような不完全な機械翻訳システムを利用しつつより円滑な異言語間コミュニケーションを実現する手法として、書き手が明示的に構造を指定し、それをそのまま読み手に見せる、という手法について提案する。さらに、そのようなコミュニケーションを支援するために開発した、インタラクティブなテキスト編集・表示システムについて紹介する。本手法において、自然言語文は、句に相当するノードと、句の間の関係を示すリンク、からなるグラフとして表現される。目的言語に変換する際には、ノード内の句をそれぞれ独立に機械翻訳にかけ、それにもとづいてリンク構造を再構成して提示する。本システムによって、これまで困難であった異言語間コミュニケーションをよりスムーズに行えるようになるものと期待できる。

1 はじめに

グローバリゼーションの進行に伴い、異言語間でのコミュニケーションの重要性が増してきている。たとえば、先の東日本大震災では、原発の状況について、外国語での情報発信が不十分であった点が批判されている。また、大学や企業においては、日本語のわからない外国人が増えているにも関わらず、大部分の情報が日本語のみで発信されている、といった問題も指摘されている。そのような状況の中、機械翻訳への期待が高まっているが、完全な全自動機械翻訳はいまだに実現されていない。単語の意味の曖昧性解消などはコーパスベースの手法によって精度が高まってきているが、係り受け関係など長い文の構造を正しく解析することはいまだに困難な問題として残されている。

そこで、本稿では、不完全な機械翻訳システムを利用しつつより円滑なコミュニケーションを実現する手法として、書き手が文の構造を明示的に指定し、さらにその構造をそのまま読み手に提示する方法を提案する。また、そのような構造が付与された文書をインタラクティブに編集、翻訳することのできるシステムを紹介する。

自然言語文は線形の文字の列として表現されるが、そもそもその文によって伝えたい概念は係り受け関係や並列関係といった構造をもっていると考えられる。通常の自然言語コミュニケーションでは、書き手がこの構造を線形の文字列に埋め込み、読み手側で埋め込まれた構造を再構成する作業が行われる。同一言語であれば、文法規則や背景知識を利用することで正しく構造が読み手側に伝えられることが期待できる。しかし、異言語間コミュニケーション、特に機械翻訳を利用した場合には、この構造を正しく伝えることは非常に困難である (図 1)。まず、書き手が頭の中にある概念を線形の文としてエンコードする際に誤りが混じる可能性がある。次に、機械翻訳器の解析部が、その線形にエンコードされた文から元の構造を取り出す際に誤りが生じる可能性がある。さらに、機械翻訳器の生成部分が、取り出した構造をまたターゲット言語で線形にエンコードする際に誤りが生じ、最後に、読み手が機械翻訳によって生成された線形の文から構造を取り出すときにまた誤りが生じる可能性がある。提案手法は、線形表現へのエンコードとデコードを行わず、構造情報を直接表現して書き手から読み手に渡すことでこのような誤りを回避しようとするものである。

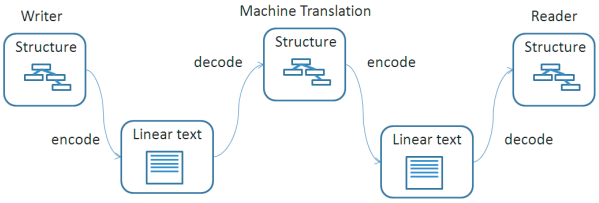


図 1. 構造をもった概念の線形文書へのエンコードとデコード

2 関連研究

機械翻訳に関する研究は数多くあるが、基本的に、通常の自然言語文を受け取って、通常の自然言語文を生成するというアプローチがほとんどである [1]。翻訳支援ツールもいろいろなものが存在しているが、やはり自然言語文を生成すること最終目的にしてお

* Takeo Igarashi, 東京大学

り、主に単語やフレーズレベルでの対訳の管理を支援するようなものがほとんどである。我々の考え方に近いものとして、文の構造を翻訳者が明示的に指定していくことで翻訳作業を支援する手法が提案されている [2]。しかし、この手法でも、構造はあくまで翻訳作業のための中間的な表現であり、最終的な出力結果は自然言語文である。また、構造は、機械が正確に処理できるようにあらかじめ用意されたテンプレートに沿って表現しなければならず、我々の提案手法のように、人間に見せることを前提とした自由な表現を利用するには不十分である。

通常の自然言語処理でも、中間処理結果として構造情報が使われている (たとえば係り受け解析など [3, 4])。提案手法は、このような構造を書き手が直接明示手的に指定して読み手に提示するところに新規性がある。既存の係り受け解析の結果をもとに書き手がそれを編集する方法も検討したが、解析に誤りが多く修正の手間がかかりすぎることと、解析の粒度が細かすぎてそのまま読み手に提示しても意味をつかむのが困難であることから、独自の表現を利用することにした。

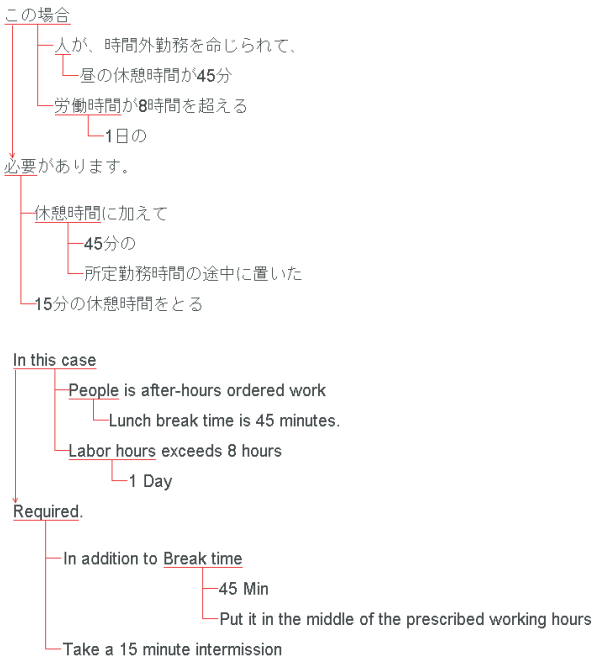
機械翻訳が不完全であることを前提に、円滑なコミュニケーションを支援しようとする研究として、制限言語を用いたアプローチがある [6]。これは、文法を厳しく制限することによって、機械にとって扱いやすい文のみを入力として受け付けるようにすることで、機械翻訳の精度を上げようとする手法である。操作マニュアルなどにおいて実用化されているが、極端に自由度が制限されるため、複雑な意味を伝えたり、日常のコミュニケーションに利用するには不自由があると考えられる。

機械翻訳の結果が正しいかどうかを、ターゲット言語を理解できない場合でも確認できるようにする手法として、折り返し翻訳による方法が提案されている [5]。これは機械翻訳によって一旦ターゲット言語に訳された文を、再び機械翻訳によって元の言語に翻訳し直すというものである。折り返し翻訳された文が正しい意味になっていれば、ターゲット言語の文もある程度正しいものと期待できる。逆に、折り返し翻訳の結果が意味不明になっていれば、ターゲット言語での文も意味不明になっていることが予想される。しかし、この手法では、結果の妥当性を確認することができるのみであり、うまくいかない場合の解決手段を与えるものではない。

計算機によるテキスト表現の拡張という意味では、関連技術としてハイパーテキストが挙げられる [8]。ただし、ハイパーテキストは、文書間の関係をリンクとして表現するものであり、文章の中の句の関係を表現する我々の提案手法とは扱う粒度が大きく異なる。また、計算機によるインタラクティブなテキスト編集支援という入力において、近年の eclipse や visual studio のような開発環境 (IDE)、日本語入力のための IME、および Microsoft WORD などに実装されているスペルチェック機能などに、ヒントを得ている。

3 提案システム

提案手法は、図 2 に示すような構造をグラフで表現したような文 (構造化文) を利用してコミュニケーションするものであり、提案システムはそれを編集するためのエディタとして動作する。書き手は、本エディタ上で、ソース言語を使って構造化文を作成する。システムは、構造つき文の中の自然言語部分をターゲット言語に置き換えて、読み手に送る。読み手は、ターゲット言語で表現された構造化文を読む。以下、まず構造化文の構成について説明し、次にそれをどのように作成・変換するかについて説明する。



原文: 昼の休憩時間が 45 分の人が、時間外勤務を命じられて 1 日の労働時間が 8 時間を超える場合、所定勤務時間の途中に置いた 45 分の休憩時間の他に 15 分の休憩時間をとる必要があります。

図 2. 構造化文の例

構造化文は、一般的なグラフと同様にノードとリンクから構成される。ノードひとつひとつが文の中の句に相当し、リンクがその句の間の関係を表現する。ここでの句には厳密な定義はなく、文の中のひとまとまりの意味を形成するまとまりであれば何でも構わない。主語と述語からなる単文でもよいし、1 つの語を修飾する長い修飾語の集まりでもよい。重要なことは、機械翻訳によって単語の羅列になってしまってもその意味を推測できる程度の、簡単な内容のまとまりにしておくことである。リンクとし

ては、単純な継承関係（Aが起きてからBが起きる、あるいはAという状況でBが起きる、など）、係り受け関係（Aを修飾する句としてのB）の２種類を提供している。継承関係の場合は句が句全体にかかり、係り受け関係の場合には句が対象の句の一部にかかっている。図２では、「この場合」から「必要があります」に繋がっているのが単純な継承関係であり、「１日の」が「労働時間」にかかっているのが係り受け関係である。リンクの場合も、ノードと同様に使い分けに厳密な規則はなく、読み手に取ってわかりやすいと思われるものを適宜選べばよい。

エディタは基本的には通常のテキストエディタと同じように動作する。すなわち、キーボードからの文字入力やマウス操作によって画面上の文字列を編集する。任意の位置への挿入や削除、コピー＆ペーストなども可能である。初期状態では、すべての入力は一つの句として与えられる。この句の上で、部分列を選択して、特定の操作（マウス右ボタンドラッグジェスチャあるいはリターンキーを押す）を行うと、その部分列が取り出されて新しい句となる。カーソルの位置およびテキストの選択状態に応じて、継承関係あるいは係り受け関係が指定される (図 3)。

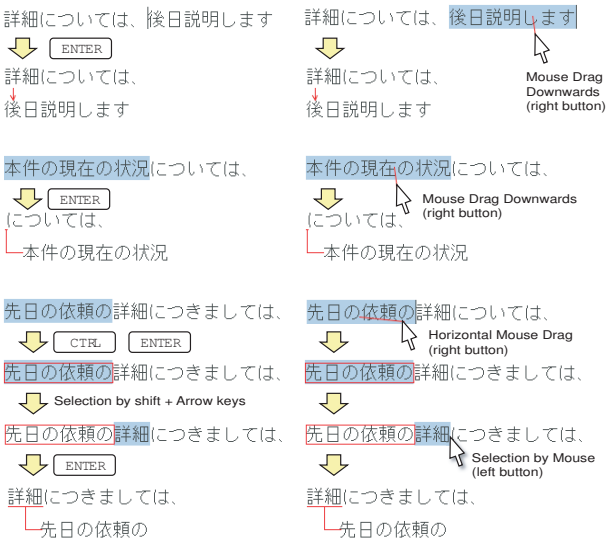


図 3. 構造化のための操作．キーボード操作（左側）およびマウス操作（右側）の両方をサポートする。上から順に、継承関係の指定、係り先のない係り受け関係の指定、係り先を持つ係り受け関係の指定。

ソース言語での構造化文を作成した後で、その構造化文のターゲット言語への置き換え作業を行う。具体的には、ノードの句をソース言語からターゲット言語に置き換える。リンク構造はそのまま保持する。出来上がったターゲット文の文書も、ソース文書と同様に、リンク構造を含めて自由に編集することができる。

4 実装

エディタは, Java Swing の JTextPane をユーザインタフェースとして利用している。内部的には、ノードとリンクからなるグラフ構造を保持しており、ユーザが JTextPane 上で文字列を編集すると、まず対応するノードのテキストを変更し、その結果を JTextPane 上に渡してレンダリングする (図 5)。レンダリングの際は、まず、参照関係に基づいてノードの表示順序を決定する。同じノードを参照するノードが複数ある場合には、参照位置が後ろのノードほど上に来るようにすることで、リンクを表す線がノードと交差することを防ぐ。次に、その順序にしたがってノードのテキストを JTextPane に渡して、その画面上の位置を得る。その際、参照しているノードがある場合には、先頭がその位置に合うように適宜スペースを挿入する。各ノードの後には改行を２つ挿入する。最後に、JTextPane によって与えられた各テキストの画面上での位置に従って、ノード間の関係を表すリンクの情報をアノテーションとして描画する。

ノード内の句の翻訳については、一般的な機械翻訳として、Web サービスとして無料で利用可能となっている Microsoft Translator API [7] を利用している。これは、日本語文をクエリとして投げると自動的に英文に変換して返してくれるというものである。ただし、変換前の文内の単語と変換後の文内の単語の間の対応関係までは、情報として返されないので、その部分を取り出すためには特別な工夫が必要である (図 4)。現在の実装では、まず、現在変換したい日本語文の中で、他の句から参照されている単語部分を、A0, A1 といった記号に置き換える。次に、置き換えたあとの日本語文を自動翻訳にかけて英文に変換する。その後、翻訳結果の英文の中の A0, A1 といった記号部分を、元の単語をそれぞれ自動翻訳にかけて英単語にしたものに置き換える。このようにすることで、翻訳後の英文のどの部分が参照されているのかを検出することができる。この方法のデメリットとして、単語が意味の無い記号に置き換えてしまうために、コーパスベースの手法を利用している機械翻訳の結果が本来よりも悪くなってしまうという点が挙げられる。また、翻訳システムをまったくのブラックボックスとして扱っているため、細かい部分を調整したり、独自の辞書を使ったりすることが難しい。将来的には、本格的な翻訳システムとより密に統合したようなシステムとして実装していきたい。

5 動作結果

図６～９に本システムを利用して作成した構造化文の例をいくつか示す。元の文は、学内の事務連絡メイリングリストに流れた内容の一部である。文の

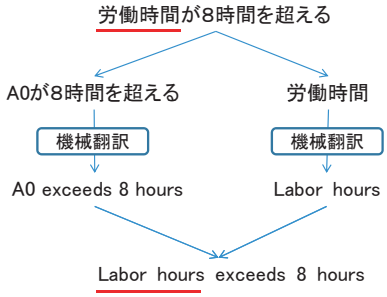


図 4. 機械翻訳機との接続

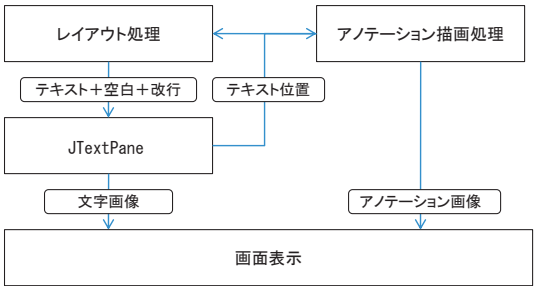


図 5. 画面表示のアルゴリズム

解釈は主観的なものなので、客観的な結論をここから導き出すことは難しいが、英語の構造化文だけを読んでもある程度は正しく意味が伝わっていることがわかる。参考までに、同じ文書をそのまま通常の機械翻訳 (Microsoft Translator) にかけた結果も併記する。通常の機械翻訳の場合には、局所的にはそれらしい表現がでてきているが、全体の構造が崩れており、それを正しく推論することが困難であることがわかる。

日本人に構造化文書を作成してもらって留学生に意味が取れるか聞いてみる実験も開始しており、ざっくりした結果として以下のような観察が得られている。それぞれの句の翻訳は通常の機械翻訳を利用しているので、その性能が結果が悪い場合には提案手法を利用しても意味を撮ることは難しい。しかし、長く複雑な文章については、構造情報を追加することで、通常の機械翻訳よりも意味が取りやすかった場合もある。ユーザがどのくらい労力をかけるとどの程度コミュニケーションがうまくいくのかについて定量的な答えを出すことは困難であると思われるが、現状のシステムを使ってみた感触としては、いずれにしても正解はなく機械翻訳によって発生する曖昧性の影響も大きいので、あまり時間をかけて作りこんでも無駄になる可能性があり、細かいことは気にせずにざっくり大局的な構造を指定するぐらいがもっとも効率がよいと考えられる。

6 議論

本手法は、不完全な機械翻訳を利用して、なんとか異言語間コミュニケーションを成立させるようとする試みであり、完全なコミュニケーションを実現するものではない。内容を１００%正しく伝えることが重要であるような文書においては、きちんと人間の翻訳者が翻訳することが必要であると考えられる。一方、本手法が有効であると考えられるのは、異言語で情報発信する必要があるが、人間の翻訳者を雇用したり書き手がターゲット言語で記述したりすることが難しく、かつ読み手に多少の不自由を強いても許容される、といった場面である。具体的には、さまざまな言語を母国語とするメンバーが存在しているような企業や大学といった組織内における、コミュニケーションツールとしての利用を想定している。組織内の場合には、共通語（英語）を使えば十分とも考えられるが、すべてのメンバーが共通語に堪能でない場合には、本手法が有効であると考えられる。ただ、本手法を使ったコミュニケーションを行う場合でも、構造化文を使った表現に慣れるための訓練がある程度の必要であり、共通語を学習する手間と比較してそれが許容範囲なのかの確認が必要である。最後に、そもそも、文章を書くのに慣れていない人や外国語の苦手な人は、文書の構造をきちんと理解していない、という問題がある。本手法を利用することによって構造をきちんと意識するようになるため、教育ツールとしての有用性もあるものと考えれる。

7 まとめと今後の課題

異言語コミュニケーションを支援する手法として、文の構造を明示的に編集・提示する方法を提案した。本手法は、書き手と読み手が協力しあうことにより、限界のある機械翻訳を利用して、プロの翻訳者を介することなく、言語間のバリアを乗り越えようとするものである。技術的な観点からは、機械の得意な作業（単語置き換え）を機械に任せ、機械にとって難しい部分（構造の解析）を人間に任せるという考え方がベースになっている。

実装上の課題としては、すでに述べたように、現状のようにブラックスボックスとして機械翻訳を用いるのでなく、処理の内容にアクセスみることのできる翻訳システムと統合することにより、本システムに最適化された翻訳処理をできるようにしていきたい。たとえば、現状では句や被参照部位単位で独立に翻訳機にかけているが、あいまい性解消などのためにはより大きな単位で処理した方が有利であるので、周りの句の内容などもコンテキスト情報として翻訳機に渡して、処理に利用することなどを行いたい。また、ユーザによる修正操作の内容から学習して、より適切な翻訳結果を返すようにもしていき

たい。また、現状のシステムは単なる孤立したテストアプリケーションでしかないので、実際の知的生産活動において使えるように、電子メイルや Wiki, 掲示版などといった、他のアプリケーションとの統合を行っていきたい。

現状でのプロトタイプシステムは日本語から英語への変換のみをサポートしているが、今後はいろいろな言語ペアに対応できるようにしていきたい。特に、日本語からベトナム語、のように、書き手が全く理解できない言語へ変換する場合に対応できるように、句や単語単位で折り返し翻訳を適用して内容を確認したり、複数の訳が考えられる場合にそれらを提示してユーザに選ばせる、といった機能を実装していきたい。さらに、本手法は３つ以上の多言語間でのコミュニケーションにも有用であると考えられるので、その方向への発展もおこなっていきたいと考えている。

提案手法の有効性を示すためには、ユーザスタディによって、提案した手法が実際にユーザに受け入れられることを確認する必要がある。すなわち、書き手にとって構造を付与するという作業がそもそも可能であるか、また、読み手にとって本システムで生成された構造化文から意味を読み取ることが可能であるのか、といった点を明らかにしていきたい。さらに、構造付与の手間および誤解釈の可能性がどのくらいか、という点も定量的に調べていきたい。

参考文献

[1] Koehn, P. 2007. Statistical Machine Translation. Cambridge University Press.

[2] 藤原 久美, 劉 紹明, 翻訳支援装置及び翻訳支援プログラム, 特開 2011-018189, 2009-07-09(出願日)

[3] KNP http://nlp.ist.i.kyoto-u.ac.jp/EN/index.php?KNP

[4] CaboCha/南瓜 http://code.google.com/p/cabocha/

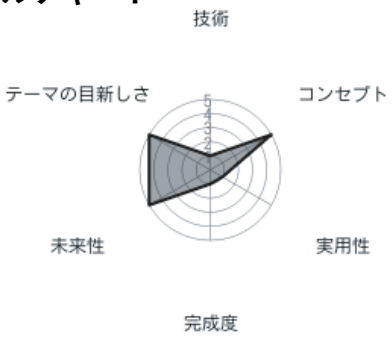
[5] Jonathan Pool, Can Controlled Languages Scale to the Web? 5th International Workshop on Controlled Language Applications, 2006.

[6] Mai Miyabe, Takashi Yoshino, Tomohiro Shigenobu: Effects of Undertaking Translation Repair using Back Translation, Proceedings of the 2009 ACM International Workshop on Intercultural Collaboration (IWIC'09), pp.33-40(2009).

[7] Jeff Conklin. 1987. Hypertext: An Introduction and Survey. Computer 20, 9 (September 1987), 17-41.

[8] http://www.microsofttranslator.com/tools/

アピールチャート

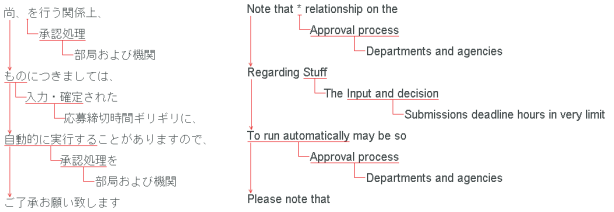


未来ビジョン

文章を表現する方法としては、文字を線形に並べる方法が唯一の方法として長らく使われてきた。しかし、文の構造とはそもそも非線形の階層構造をなすものであり、本論文で提案するようなグラフィカルは方法は、より自然な方法であると考えられる。実際に、外国語を新たに勉強するための教科書などでは、文章の係り受けや入れ子構造などをグラフィカルに表現することが行われている。

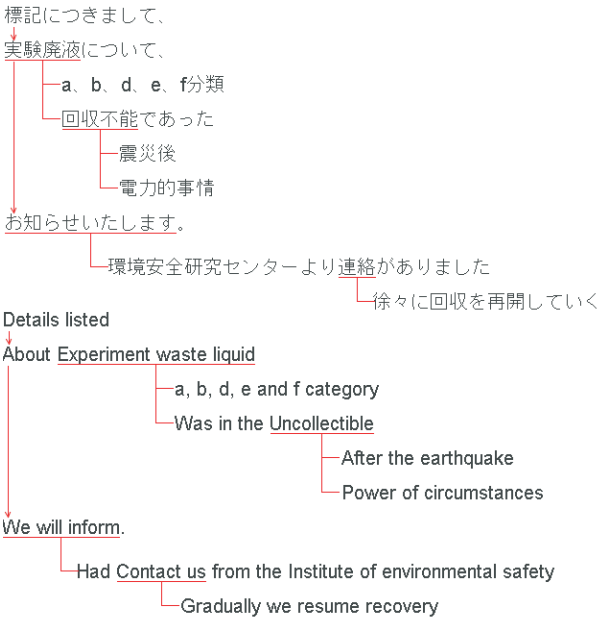
グラフィカルな表現が日常の文章表現に使われない理由としては、編集の手間がかかることと、スペースを余分につかってしまうことなどが考えられる。しかし、これらは計算機システムを使うことでかなりの部分解決できるのではいかと期待できる。本稿で提案したシステムは、そのような考えに基づいた試みの第一歩である。将来的には、計算機システムを活用したインタラクティブな手法を開発していくことにより、グラフィカルな表現を活用した、よりゆたかな自然言語表現および言語を介したコミュニケーション手法を提案していきたいと考えている。

また、本研究の背景にある問題意識として、現状の機械翻訳システムがブラックボックスとして提供されているという点があげられる。そのため、翻訳結果において別の訳語を選ばせる、といった程度のことはできるが、構造の解釈が間違っている場合にそれを手作業で部分的に修正するといったことはできないようになってしまっている。本研究では、このようなブラックボックスとしての機械翻訳の外側を工夫することで問題に対処しようとしているが、今後は、ユーザインタフェースを工夫することによって、ブラックボックスの中身に直接アクセスできるような仕組みも提案していきたい。



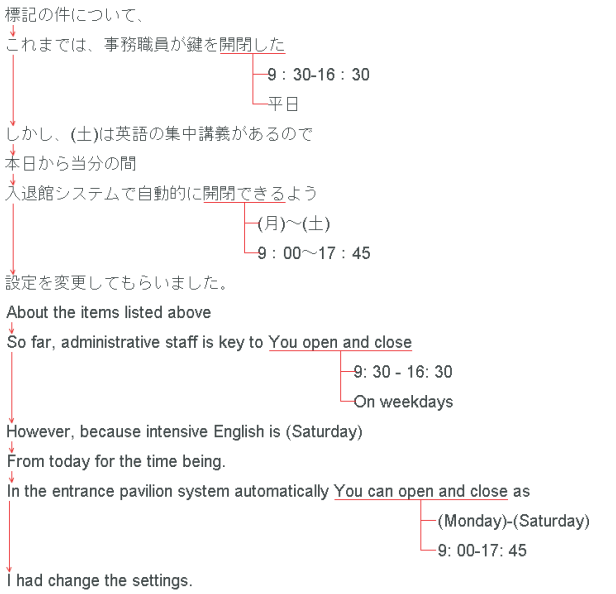
尚、部局および機関承認処理を行う関係上、応募締切時間ギリギリに、入力・確定されたものにつきましては、部局および機関承認処理を自動的に実行することがありますので、ご了承お願い致します。
Do the departments and agency approval process relations, submissions deadline time last-minute onto, we kindly ask regarding input, deterministic, one departmental and agency approval process to run automatically, so please understand that.

図 6. 結果の例 1



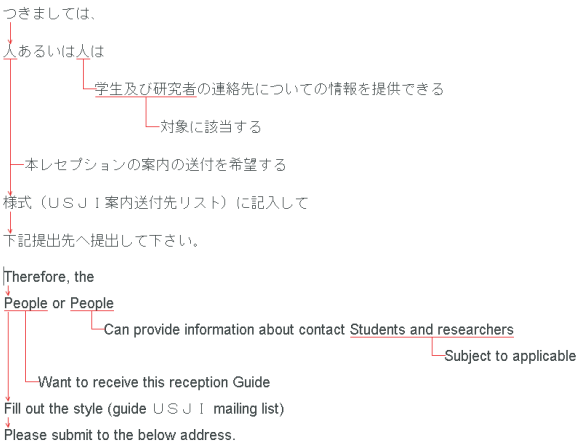
標記につきまして、震災後電力的事情で回収不能であった、a、b、d、e、f 分類の実験廃液について、徐々に回収を再開していく旨環境安全研究センターより連絡がありましたのでお知らせいたします。
Inform stating that listed details about the experiment was, in the circumstances of power uncollectible in after the quake, a, b, d, e, f classification of waste water will resume recovery gradually from the Institute of environmental safety had contacted.

図 7. 結果の例



標記の件について、これまでは、平日の 9：30-16：30 事務職員が鍵を開閉して来ましたが、(土) は英語の集中講義があるので、本日から当分の間、(月)～(土) は自動的に 9：00～17：45 の間、入退館システムで開閉できるよう設定を変更してもらいました。
Matters listed so far, weekdays 9: 30 - 16: 30 I change settings and office workers came closes key is (Saturday) is because intensive English, from today for the time being, is (Monday)-(Saturday) between 9: 00-17: 45, in the entrance pavilion system closes automatically can be.

図 8. 結果の例



つきましては、本レセプションの案内の送付を希望される方、あるいは対象に該当する学生及び研究者の連絡先についての情報をご提供いただける方は様式（ＵＳＪＩ案内送付先リスト）に記入の上、下記提出先までご提出下さい。
You still want this reception information, or criteria for students and researchers contact information you'd like to contribute towards a style (guide ＵＳＪＩ mailing list) filling up below submit to must submit.

図 9. 結果の例