CICP2012 プロジェクト提案書（※提案書全体で4ページ以内を厳守のこと）

1. プロジェクト名

コーパスが、君の論文の英語は変だって言っていたよ。

2. プロジェクトリーダー（※同一学生が複数プロジェクトのリーダーにはなれません）

所属研究室	学年	学生番号	氏名	e-mail アドレス
自然言語処理学研究室	M2	11510054	澤井悠	yu-s@is.naist.jp

3. 分担者（※他大学の学生も可．分担者無しも可）

所属研究室	学年	学生番号	氏名	e-mail アドレス
自然言語処理学研究室	D2	1161009	林部祐太	yuta-h@is.naist.jp
自然言語処理学研究室	M2	1151052	坂口慶祐	keisuke-sa@is.naist.jp

4. チューター（※必須．本研究科専任教員，伝票入手完了までサポートする義務あり）

所属研究室	職名	氏名	e-mail アドレス
自然言語処理学研究室	助教	Kevin Duh	kevinduh@is.naist.jp

5. 必要経費（※応募時点での見積書添付は不要）

	金額（千円）	支出予定月	品名・型名・数量／行先・目的・日数等
設備備品費 ※支出は 12 月末まで、以降は支出不可の場合がある ※全体の 70%を超える場合、本表空欄に理由を明記	300	7 月	開発用コンピュータ（Precision T1650）2台
消耗品費 ※通常の研究費でも購入可能な物品に限る	80	7 月	ディスプレイ（DELL U2410，24 インチ）2 台
	30	7 月	キーボード・マウス等の周辺機器
	200	7 月	英語教育，プログラミング，言語処理等に関する書籍 50 冊
	50	7 月	Microsoft Office（アカデミック版）3 個
	30	7 月	Adobe Acrobat （アカデミック版） 2 個
	30	7 月	Microsoft Windows 7 Home Premium 2 個
	80	7 月	HDD（SEAGATE ST3000DM001）5 個
旅費（調査目的も可） ※国内・海外いずれも可 ※交通費+宿泊費（実費）のみ ※日当・人件費・謝金は不可	200	12 月	国際会議 COLING（インド） 調査，1 週間， 1 人
合計（上限 1,000 千円）	1000		

6. プロジェクトの背景と目的

【背景】英語を外国語として用いる人々(EFL)は 10 億人を超えている[1]. そのため, これまで EFL の作文から, スペル誤りや冠詞・前置詞等の文法誤りを自動的に取り除く研究が盛んに取り組まれてきた. 一方, 文法誤りから一歩進んで,「ネイティブはあまり使わない表現をより自然な表現に修正する」といった**意味に関する誤りの自動検出**はあまり行われてこなかった. その主な理由は次のとおりである.

(1) EFL がテーマを限定せずに自由に書いた文章(**自由作文**)には文法誤りが多くあるため, 意味誤り検出を行うのに必要である構文解析等が困難である

(2) 自由作文では様々な話題があり, 用いられる語彙が膨大である. そのため, 多義語の曖昧性の解消が困難である.

【目的】本プロジェクトでは, 前述した問題点を回避して, 意味誤り検出に取り組むため, 本プロジェクトでは**対象とする文書を科学論文に絞る**. **科学論文には**,

(1)執筆者がネイティブでなくても基本的に高等教育を受けているため, 一般的な EFL と比較して英語能力が高い. そのため, 自由作文と比較して文法誤りが非常に少ない.

(2) 話題が限定されているため多義性のある語彙も少ない

という特徴があり, 意味誤り検出に取り組むのにふさわしい対象であると考える.

EFL が論文を執筆する際, 問題となるのは, その学術分野特有の英語表現を使いこなすことである[2]. そこで, 本プロジェクトでは, オープンアクセスの論文集から, ネイティブの文章と EFL の文章を大量に収集し, ネイティブがよく使う表現と, それに対応する EFL が(誤って)よく使う表現の対を統計的に獲得する. そして, EFL が入力した表現に対して誤りがあれば, より"自然"な表現を修正候補として提示するシステムを作成する.

[1]http://www.britishcouncil.org/learning-elt-future.pdf

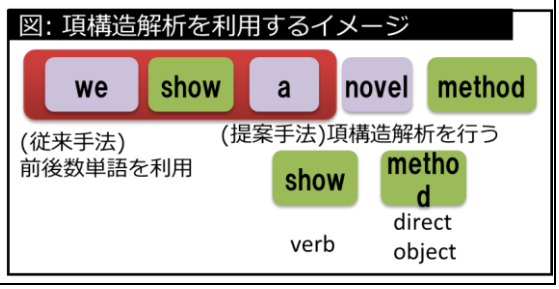
7. プロジェクトの独創性・特長

以下の優遇条件に該当する場合はチェックボックスに印を入れること:

□環境テーマ, □安全安心テーマ, □福祉テーマ, □多国籍チーム

【特長1】これまで, EFL が間違いやすい表現は, 主に英語教師の経験に基づいてまとめられてきた. しかし, この方法は専門的な文章を理解できる英語教師は少ないこと, 人の目では見落としが発生すること, が問題点である. そこで本プロジェクトでは, 統計手法を用いてネイティブの文章と EFL の文章から自動的に, そのような表現を抽出することで対処する点が特長である.

【特長2】従来研究の誤り訂正の研究では, 検査する単語の前後数単語(n-gram)を手掛かりに誤り検出が行われてきた. 本プロジェクトは**述語項構造を手掛かりに**して検査を行う. 述語項構造とは, 各単語の位置や修飾語の有無に依存せずに, 述語に関係する主語や目的語(項)をとらえることができる構造である. そのため, 長い文章から適切に誤りを訂正するのに必要な手掛かりを適切に取り出せると考える.



8. 目的到達までの研究計画

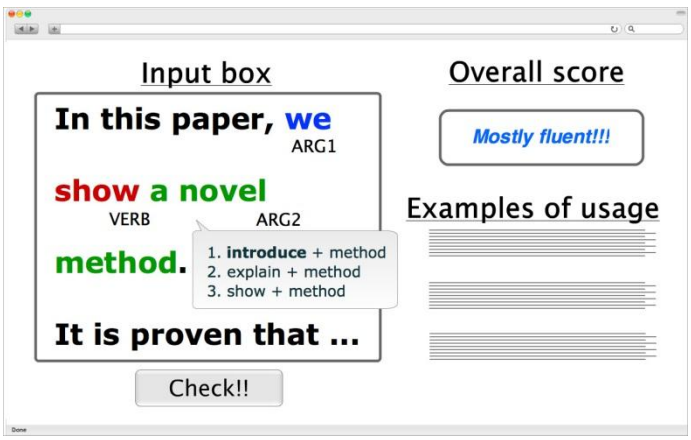
【構築するシステムの概要】

右図は本プロジェクトで作成するシステムのイメージである.

文章が入力されると, システムはネイティブがあまり使わない不自然な箇所を検出し, その箇所をハイライトする. そこをクリックすると修正候補の提示を行う.

本プロジェクトでの誤りを検出する箇所は述語とその項に絞る. 右の例では動詞 show の代わりに introduce や explain がネイティブはよく使うことを示している.

将来的には不自然な形容詞や動詞の用法の検出にも取り組みたいと考えている.



【予想される困難と回避方法・代替案】

(1)入力にスペル誤りが含まれている可能性がある

ノンネイティブの文章には，スペル誤りが混入しやすい．スペル誤りが検査するフレーズの中に入っていると，頻度の比較計算に支障をきたす可能性が高い．そのため，事前にできる限りスペル誤りは修正しておくべきである．そのため，本システムが処理する前に GNU Aspell 等のスペル訂正ソフトウェアを用いる．

(2)筆者の意図を汲み取ることは非常に難しい

本プロジェクトでは筆者が書いた文が自然であるか否かの判定を行い，自動的な訂正は行わない．これは「訂正」までしようとすると，筆者が何を伝えたいかという意図を汲み取る必要があるが，現在の言語処理技術ではそれは困難であるためである．

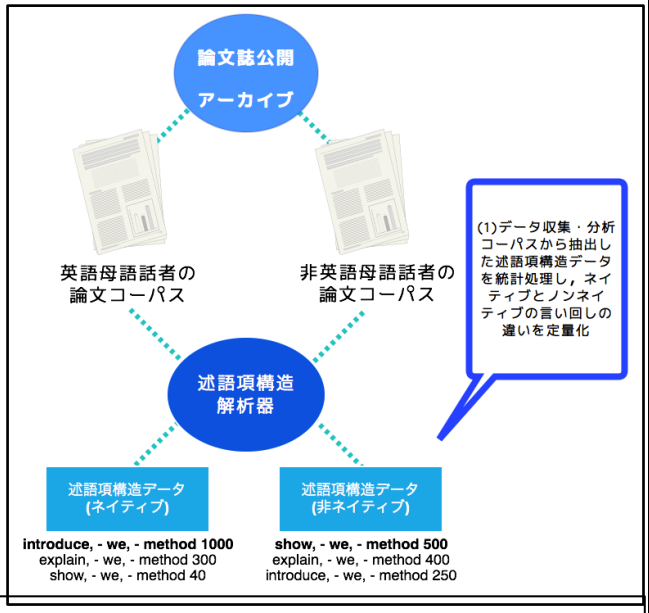
【研究計画】

(6 月～7 月)コーパスの構築と前処理

ネイティブとノンネイティブのコーパス(文書集合)を構築するために，誰でも自由に閲覧できる論文集サイト ACL Anthology を用いる，ここでは，論文が PDF 形式で公開されており，さらに著者情報も BibTeX 形式で公開されている．ここから論文を大量に取得する．

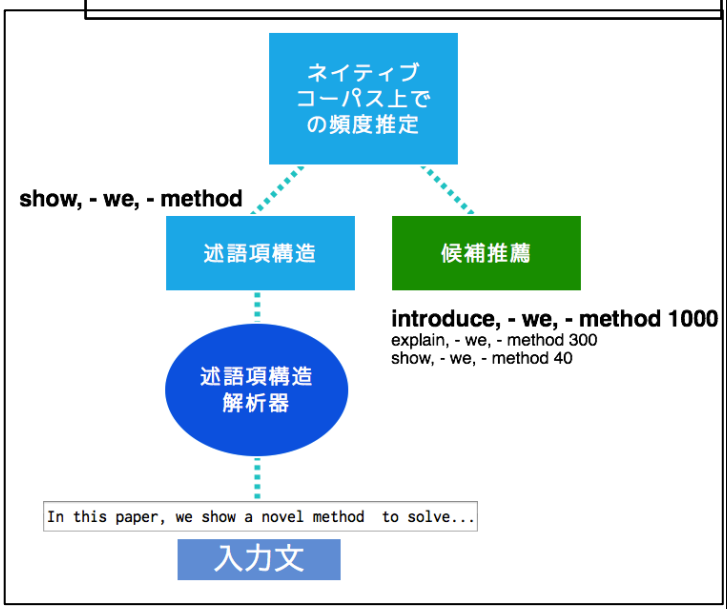
そして取得した PDF から PDF2Text を用いてプレーンテキストに変換する．そして，項構造解析を行うために，Stanford Parser 等の解析ソフトウェアを用いて，品詞付与・構文解析を行う．

次に， BibTeX 中の第一著者のメールアドレスや人名を手掛かりにして，各文書の著者の母語を推定し，ネイティブと EFL の文書を区別する．



(8 月～9 月)システムのプロトタイプとテストセットの作成

ネイティブと EFL のコーパスから統計手法を用いて，ネイティブがよく使う表現と，EFL がよく使う表現の対を抽出する．また作成したシステムの性能評価を行うため，「不自然な表現」が含まれた文をノンネイティブコーパスから取り出して，テストセットを作成し，適宜性能評価を行う．模擬国際会議では，作成した言語資源を分析して得られたノンネイティブが犯しやすい誤りについて報告する．また，プロトタイプの構成と性能についても報告する．



(10～12 月)改良・ローカル環境でのシステム構築

プロトタイプを用いて得られた予備実験の結果から、システムが不自然な箇所を発見できなかったり、修正候補を提示できなかった事例を分析する。その結果を踏まえ、手法を発展させてさらに高性能に不自然な箇所を検出できるようにする．また，コマンドライン上で任意の文書に対してシステムが動作できるようにする

(1 月～2 月)他分野への応用および公開用ウェブインタフェースの構築

一般公開用のウェブインタフェースを構築する．そして，使い勝手や細かなシステムの改良を行っていく．また，ある分野の論文の PDF さえ用意すれば，自動的にその分野の論文に対応できるようにする

(3 月)ウェブシステムを学外にも公開し，フィードバックを得る．

9. 学位論文との関連・相違点(※複数メンバによる実施の場合はメンバごとに記述)

澤井は修士論文では，EFLの作文中の動詞選択の誤りの訂正をテーマとする予定である．EFLの意味に関する誤りの訂正という点は本プロジェクトと共通するが，動詞のみに注目する点と，対象とする文章が論文ではなく様々な話題を含む自由作文とする点で異なる．

林部は博士論文では，修士課程で研究してきた「述語項構造解析技術」を応用して，EFLの作文中の前置詞誤りの検出と訂正に取り組む予定である．述語項構造解析とは，述語に対する主語や目的語(項とよぶ)を自動で抽出させる処理である．用いる前置詞の決定には述語と項が何であるかが大きな役割を担うため，述語項構造解析が利用できると考えた．本プロジェクトとは述語項構造を用いる点が共通しているが，修正対象が前置詞である点と，対象とする文章が英語学習者(高校生程度)によって書かれた話題を限定しない英作文ではなく，ある程度英語が出来る研究者によって書かれた特定分野の話題に限定した文章であるという点が異なる．

坂口は修士論文において「英語学習者作文に対する頑健なスペリング誤り訂正」に取り組む．本プロジェクトとは，英語学習者作文における誤りを検出する点では関連している．しかし，スペリング訂正が形態論的(表層的)な誤りを訂正するのに対して，本プロジェクトでは意味論的(深層的)な誤りを訂正することを目指している点が異なる．従って修士論文と本プロジェクトのテーマをあわせて行うことで，相互補完的な役割を果たすものと期待できる．

10. 研究業績(※複数メンバによる実施の場合はメンバごとに記述)

澤井:1，林部:2～5，坂口:5～6

1. Yu Sawai, Xiaodong Niu, Florian DeVuyst, Hiroshi Yamaguchi, Measurement of concentration in solid-liquid two-phase flow using magnetic fluid, Physics Procedia, Volume 9, 2010, pp. 137-141(査読有)
2. Yuta Hayashibe, Mamoru Komachi and Yuji Matsumoto , Japanese Predicate Argument Structure Analysis Exploiting Argument Position and Type, IJCNLP, 2011, pp.201-209 (査読有)
3. 林部祐太, 小町守, 松本裕治, 隅田飛鳥. 「日本語テキストに対する述語語義と意味役割のアノテーション」, 言語処理学会第18回年次大会, 2012, pp.1-4 (査読無)
4. 林部祐太, 小町守, 松本裕治. 「文脈情報と格構造の類似度を用いた日本語文間述語項構造解析」, 情報処理学会第201回自然言語処理研究会, 2011,No.10, pp.1-8 (学生奨励賞受賞) (査読無)
5. Keisuke Sakaguchi, Yuta Hayashibe, Shuhei Kondo, Lis Kanashiro, Tomoya Mizumoto, Mamoru Komachi and Yuji Matsumoto, NAIST at the HOO 2012 Shared Task, The 7th Workshop on Innovative Use of NLP for Building Educational Applications, 2012 (査読無)
6. 坂口慶祐, 水本智也, 小町守, 松本裕治. 「英語スペリング訂正と品詞タグ付けの結合学習」, 情報処理学会 第206 回自然言語処理研究会, 2012 (査読無)

11. チューターから一言

I think this is a very promising project.

First, it is a good idea with practical impact. I would love to use the system at NAIST.

Second, it incorporates interesting research elements, such as predicate argument structure analysis, which sets it apart from existing work.

Third, the three student members, in my opinion, form a very strong team and I am confident that the project will be completed successfully. In sum, I am quite excited about it.