作文指導のための作文添削データベースの構築

A Japanese Revision Example Database for Essay Instruction

竹内 和広

Kazuhiro TAKEUCHI

大阪電気通信大学 情報通信工学部, 寝屋川市初町 18-8

Osaka Electro-Communication University, 18-8 Hatsucho, Neyagawa, Osaka

あらまし:現在の言語処理の技術では計算機を用いた自動的な作文指導は難しい。言語処理において、処理対象の文例のデータベース(コーパス)は必須の存在である。しかし、言語処理の技術開発に適合するように、作文添削例の事例記述データベースを開発するためには、複数の教育者の意見を統合することが前提となり、その要件については必ずしも明らかではない。本稿では、日本語母語話者が記述した作文の特質を紹介し、その特質を特徴化する言語処理上の問題点を議論する。さらに、その議論を踏まえ、協調的な作文添削データベースの構築を提案する。

Abstract: Limited by the present technology, the automatic essay instruction is not an easy task in the area of Natural Language Processing (NLP). A database (corpus) that contains language use examples is indispensable for the development of a new NLP module. However, the practical construction of the database in order to accommodate revision examples into the technical development premises a pedagogical consensus of which details are not discussed well. In this paper, we first point out some specific problems of Japanese essays written by Japanese native speakers. We then propose our ideas for building a collaborative essay instructive database.

キーワード:自然言語処理, 作文添削, 協調データベース構築

Keywords: NLP, automatic essay instruction, collaborative database development

1. はじめに

近年、小学から大学教育に至るすべての教育現場において、学生の国語力低下が嘆かれて久しい。また、インターネットの一般家庭への普及や携帯電話でのメールによるやり取りの日常化は、学生が主体的に文章を書く機会の増加につながったものの、話し言葉に近いくだけた表現を、当たり前のように作文で用いてしまう弊害も生んでいる。このような背景は、作文指導に対する切実な需要を顕在化させるに至っている。

作文指導は、指導される側の読書量、作文量といった個別の経験的要因に左右される部分が大きいため、個別指導が適している。高校あるいは大学においての作文教育は、学生が公的に通用する文章を書く能力が習得済みであることを前提とした、意味内容に踏み込んだ指導が本質であるべきだが、上記のような背景から、書き誤りの訂正や、記述方法の初歩的な誤りといった表層的なレベルでの指導も必要となる。そのため、結果として、作文指導を行う上での教育コストは必然的に高いものとなる。

また、作文指導は、習得された教員の技能・経験によるところが大きい。初・中級の外国語教育とは異なり、日本語母語話者に対する作文指導では、教員の言語直観から作文の問題を指摘する局面が多くなる。このような言語直観は教員個人に依存する特性であるため、指導される学生側から見た場合、長期間一貫した教員に指導を受けることが理想的である。しかし、教員が特定の学生に対して個別指導を一貫して長く続けることは教育コストの問題から現実的ではない。

本シンポジウムでの本発表は、教員が日々行う作文の添削のうち、言語表層レベルの誤りの添削を形式化し、データベースとして蓄積することを提案する。具体的には、以下のような特徴をもつ Collaborative な作文添削データベースの構築・運用について議論したい。

- 教師間で情報を共有する機構を用意
- 教師の添削は言語データに対するタグ付け(アノテーション)として扱う

教員が内容に踏み込んだ指導に集中するためには、学生が無頓着に使用する表現が、口語的であるのか、分かりにくいのか等、知的な作文の表現として不適切であるのかを、自分自身で確認することが望ましい。提案するデータベースは、このような学生の自学自習に利用できる、表現レベルに対しての自動作文添削ツールの構築に貢献することを中・短期的な目標としたい。

2. 添削事例蓄積のための自然言語データベース

2.1 自動作文添削と評価基準

IEEE の委員会による議論[1]にも見られるように、作文を自動採点する挑戦は現在に至っても言語処理の重要な課題である。作文の自動採点では主に教育の分野で検討がなされてきた評価基準を文章特徴として導入する。例えば、誤字脱字、表現の自然さ、論理性といった特徴である。しかし、このような特徴の詳細を精緻に、現在の自然言語処理の基礎技術では分析することは難しい。

計算機による自動添削が困難である理由は、それが自然言語理解に繋がる問題だからである。作文を添削する上で、書かれた「内容」が最も重要な観点であることは間違ないが、このような「自然言語理解」に関わる人工知能的課題は、到達する上での中間課題でさえ、定義することが困難であるといえる。

作文の自動採点が採用する研究の方向性には、記述内容の「良さ」に対する近似を、語の分布によって表現しようとする試みも多い。しかし、本稿は単なる作文の総合的な採点を目的とした議論をするわけではない。本稿が提案する添削データベースの構築は、人間が添削した事例を蓄積し、その中から、有益な言語分析に繋がる基礎的なモジュールを段階的に構築していくスタンスに基づいている。言語処理の基礎モジュールは、ともすれば、応用範囲が一般的すぎるため、目的達成を評価することが難しい。しかし、添削という人間の言語処理のしくみが解明できるのであれば、それは一般的かつ応用にも即した合理的な言語処理の基礎モジュールの提案にも役立つと考える。

2.2 コーパス

言語学・言語処理の分野では、コーパスとよばれる文データベースを研究の基盤的資源として研究をすることが多い。コーパスとは、自然言語の文を収集した一種の文例集である。

外国語教育の分野ではコーパスと計算機を利用した辞書作成が行われてきた。例えば、Collins の出版する COBUILD 英語辞典[1]はその作成のために、語が実際に使われた、小説、新聞、雑誌などの生きた文例を大量に収集し、コンピュータを援用した多角的な統計分析により、各索引語への記載項目を検討したものである。具体的には、コーパスを利用し、特定の語が、どのような語と共起するか、どのような文脈に出現するかを統計的に検討することができる。例えば、動詞"make"の右側文脈には、1語範囲内に"～ing"が出現する例がどの程度出現するか、といったことが定量的に検討可能となる。これは、人間が語の用法を現実の使用に基づいて知ることができる方法として非常に有益である。現在このような辞書作成方法は、COBUILD に限らず、多くの辞書で採用されており、辞書編纂のためには、大規模なコーパスが必須となってきている。このように、コーパスを準備し、語の共起情報を始めとする統計情報に基づいて語の用法を記述する方法論は、統計情報の利用の仕方、および、その用法の記述方法は、辞書編纂方針の本質的議論であり、辞書編纂者の辞書編纂の経験・知恵によっている。

2.3 情報タグ付きコーパス

言語処理の分野では、文例を収集しただけのコーパスに対し、コーパスに何らかの付加情報をつけたコーパスをタグ付きコーパスと呼び、言語処理モジュールを開発する上での重要な研究資源となってきた。

本稿が議論する作文添削データベースも、短期的には、表現的な誤りを書き手自身に指摘する言語処理モジュールを開発することを目的として、このような処理モジュールを開発するためには、どのような付加情報がコーパスに必要かを議論する。

タグ付きコーパスの中で言語処理モジュールの開発に貢献してきたものに、品詞タグ付きコーパスがある。自然言語は単なる文字列ではなく、規則性をもった文字列であるが、その統語・意味的な性質を記述するための基礎情報として一定の共通認識が得られているのが品詞タグであり、具体的には、文字列の中で、語と認定する文字列範囲と、その語が分類される品詞を多くの場合人手により、分析、情報付与したコーパスである。

品詞タグ付きコーパスの他にも、係り受け情報付与、固有表現情報付与等の高次の言語情報をタグ付けした言語処理技術開発向けのコーパスは、近年、数多く構築がなされて、言語解析モジュールの開発に役立てられている。本稿が目標とするデータベースの構築

---

1 http://www.collins.co.uk/

は、中・短期的には自動作文添削の実現を視野にいれており、それを実現するための言語解析モジュールを開発するための基盤となる作文添削データベースでなくてはならない。

添削をタグ付けとしてコーパスに情報付与する試みに類似する先行的な試みとして、外国語教育を目的に作成された誤りタグ付きコーパスは、興味深い。例えば、NICT JLE コーパス[2]では、英語学習者が行った誤りが、図1のように記述される。図 1 では、例えば、"He like dogs."という誤り付きの文に対して、誤りの指摘と訂正は、角カッコで囲まれたタグで情報付与される(実際の NICT JLE のタグ体系は例で示したよりも精緻であるが、紹介のため、タグを日本語化し、かつ簡略した例を掲載した)。

誤りタグを利用すると、例えば、学習進度が異なる学習者の誤りの種類に関する分布を比較し、学習初級者がよく犯す誤り、中級学習者が犯しやすい誤りといった検討ができるようになる。また、誤りタグに基づいて、特定の誤りを自動指摘する言語処理モジュールも開発されるようになっている。

しかし、このような誤りタグ利用法の前提は、誤りを一定の基準下で認定することであり、誤りタグは現状では人手でタグ情報を付与しなくてはならない以上、タグ情報を付与するためのマニュアルを作成しなくてはならない。そのため、誤りタグ付きコーパスは、誤りの指摘・訂正のマニュアル化がしやすい外国語学習における学習文法的な誤り体系をあらかじめ設定することになる。このことは、タグ付け作業者が学習文法に精通し、かつ、タグ付けマニュアルを熟知していることを必要とし、言語学や言語教育を専門としない複数の教員が協調して添削データベースを構築する方法としては課題を残すことになる。また、学習文法の体系は、日本人が日本語で作文をした時に犯す微妙な「誤り」を分析するには向いておらず、我々が目的とする添削では、あらかじめ誤りの体系を仮定することは難い。

```
He <ERR 誤り種類= "主語との一致" 誤りの訂正="likes">like</ERR > dogs.
She is active in <ERR 誤り種類="冠詞に関する誤り" 誤りの訂正="the"> a </ERR>
development of low cost water pumps.
```

図 1. 誤りタグの例

他方、厳密な誤りタグ体系を設けずに、学習者コーパスをタグ付きコーパス化している事例もある。名古屋大の日本語学習者コーパス[3]では、誤り例とその訂正を指摘するだけでなく、誤りをタグで指摘する代わりに、誤用の分析を自然言語で指摘する方式をとっている。自然言語による誤りの分析(理由)の記述は、添削を多様に記述できるメリットでもある反面、この記述の多様性が、誤りパターンを記述から特定することを困難にし、再利用の際に問題をきたす。そのため、例えば名古屋大日本語学習者コーパスでは、記述に使用するキーワードをあらかじめ決めておくなどの工夫がなされている。

このように、誤りタグの設計・体系化をした後、コーパスを作成する方向性は、学習者が誤りを犯す類型をあらかじめ体系化して、誤り部分および訂正案を制約できるという点で利点があり、他方、誤り分析を自然言語で記述する方向性は、学習者の誤りの多様性を記述できる可能性をもつと言える。

本稿が対象とする、高等教育における作文指導で扱う誤りは、学習文法の範疇でとらえ切れるものではなく、コーパスの構築と共に一般化していく必要があるため、後者の方向性が参考になる。そこで、本稿は、後者の考えを発展させ、添削データベースの構築に関して、信頼性の高さと、協調的な構築を目的とした記述方法の検討を行う。

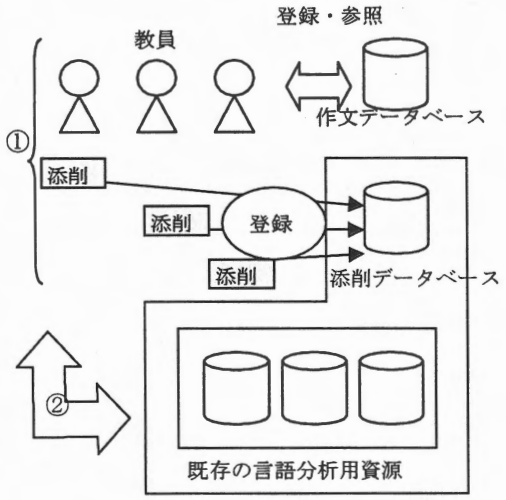

図 2. 協調添削環境

3. 協調タグ付け環境

ここまでに説明したような背景から、我々は作文添削データベースを複数の教員の協調タスクとして構築することを提案する。図2にその概念図を示す。教員は、作文事例を参照し、図中の①のように、添削を文章へのタグ付け(アノテーション)の形で添削データベースに登録する。図中の②は添削をするための支援を示す。

インターネットの普及により、協調タスクの環境は、当たり前に行われるようになってきた。例えば、インターネット上のWikiPedia[2]は、協調的に多言語の辞典を作り上げた例である。WikiPediaの基盤となったWikiという協調編集システムは、複数人がWebページを構築できるシステムである。Wikiは協調して作成しているWebページの追加・削除・変更といった編集が協調的に行う、ある種のグループウェアともいえる。

本稿が提案する作文添削データベースは、このようなネットワーク環境での、協調的な編集を提供するシステムである。本提案で、教員が協調的に追加していく本質的なレコードは添削事例となるため、図2中の②の支援は、協調的に作業するために以下の3点の要件を満たすことが必要であろう。

要件 1. 他の教員が類似の事例に対してどのような添削をしたかを検索できる。
要件 2. 添削した例、理由記述が適切に行える。
要件 3. 添削する際、客観的根拠を参照できる

また、添削は、ひとつの添削事例につき、次の組を記述することによって行う。

X:誤り範囲
Y:書き換え例(正例)
Z:誤用の記述(なぜXを誤用と判断したか)

協調的に添削事例を増やしていく試みであるので、1つの文例に対して、いくつも添削があっても構わない。また、添削についてのコメントも自然言語で記述可とすると添削に関しての議論のために有益であろう。ただし、添削のX,Y,Zの記述には、キーワードの統一よりもさらに制約の厳しい記述方法を用いたい。詳細は、次節で議論するが、そういった方法を検討する主眼は、添削という言語操作をできるだけ普遍的に記述する道具立てを視座にしているからである。

4. 添削の形式的記述

4.1 日本語話者作文の添削

外国語教育における言語指導と、母国語教育の言語指導における、不適切な作文の添削指導では、問題点が異なる。より良い日本語の指導が、外国語教育と異なる例として、「誤り」が言語直観に近づく、というものがある。

例えば、日本人がより母国語である日本語能力をより向上させることを目的として書かれた大野[4]によれば、次のようなという適切な日本語ではない例が挙げられている。

「任せておけば、しっかりしている。」[3]

そして、大野は、このような文を不適切と感じるためには、日本人が母語である日本語に対しての感覚を鋭くして行く必要があると述べている。しかし、このような、一般的な学習文法では扱えないような繊細な意味の相違は、現在のところ計算機によって計算することも実用化に至っていない。そのため、自動添削も現状では実現できない。これは、現在の自然言語処理技術では、くだけた表現や、誤りを含む表現を言語解析する点では弱く、多様な言語運用データを処理するには柔軟性に欠けることを示しており、添削事例を蓄積・分析することは自然言語処理上の一般的な課題ともいえる。

本稿が提案する方向性は、作文添削をデータベースに蓄積し、頻出する添削指導については、段階的に自動的に添削できるよう、自然言語処理の解析モジュールの開発および、性能を向上していくことである。

4.2 誤用記述の形式化

本稿が提案する添削3項組は、自然言語処理の解析モジュールと整合性がある記述である。これは、誤りとその添削後の正例を、言語処理上の計算として定義する狙いがある。

基本的なアイディアを示す。例えば、誤字のレベルの間違いは、Xの文節列とYの文節列、そしてZは単に誤字と記述して、図3のように記述する。

このような記述を蓄積すれば、添削理由が「意外」の書き誤り、であることは、データベース蓄積後にXとYの間の編集距離を参考に知ることは容易であろう。また、この例のような文節の認定は、現在の自然言語処理においても、相対的に高い精度で分析できるため、教員は、タグ付けのため、書き換えに適当な誤りの範囲を指定することが基本である。誤りの書き換え範囲を明示的に与えることは、誤りの自動判定をする上で有益な情報となる。

さらに複雑な誤りに対しては、より高度な添削をする上では範囲指定の特徴付けだけでは不十分な点もある。そこで、現在、多様な言語情報のうち、自動的には解析できないが、将来有益な基本的な言語解析モジ

---

2 http://wikipedia.org/

3 大野[]中 p.12 の練習問題より引用

ュールになるであろうものを、人手で分析し、添削事例に記述する詳細化を提案する。

多様な言語分析情報のうち、まず、思い浮かぶのは統語情報であろう。自然言語処理では、精緻な機械処理文法の入力表現として、日本語の統語構造を簡略した係り受け解析を前提とすることが多く、新聞記事などの文に関しては、係り受け解析は一定の精度で解析が可能となっている。残念ながら学生が書く自由作文に対して、このような言語解析ツールによる解析結果は必ずしも高い信頼性を見込めないこともあり、学生が書く特徴的な「誤った係り関係」を明示することは、非常に重要な情報となる。

図 4 に係り受け情報を使った添削の例を示す。

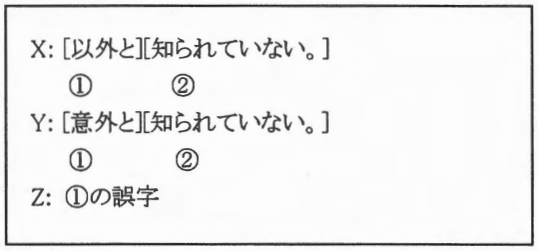

図 3. 誤字の添削例

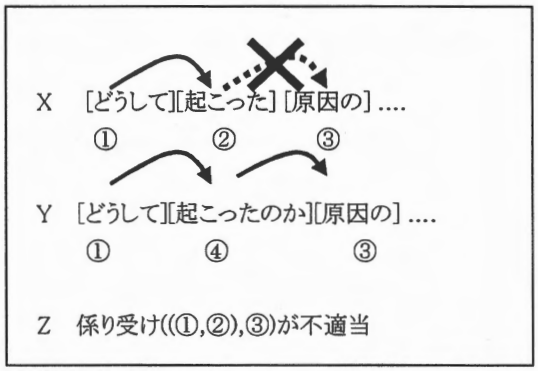

図 4.添削記述の X,Y に表現特徴情報(係り受け)を付与した例

図 4 で示したように、Z の誤用記述の問題を「副詞「どうして」の呼応の誤り」や、「どうして起こった」と「原因」の整合性がない」と書くのではなく、具体的な対象をタグで囲み、対象との相関により誤りをタグ付けする。また、前述のように、不自然な日本語は、言語解析でも不適切な解析結果になることが多いため、例えば、X の②と③の係り受けが不適切であることを、教育者が明示的に、誤った係り関係であると指摘する。

ところで、上記までの特徴記述において、便宜上、記述対象の最小単位を文節としたが、日本語の文末の構造は文節よりも大きな単位で扱った方が合理的な場合が多い。精密な議論は省くが、複合的な超文節を文末に限っては類型的に表現することが適切であることを補足しておく。

4.3 高次の用法添削の記述

教員にとって、誤用の認定は、必ずしも高い確信を持って判断できるものばかりではない。一般的に文章の良さの評価は大きくわけて2つの側面から評価されることが多い。1つは 4.2 節で提案したような、言語的装置で扱える対象から特徴づけする側面、他方は、それ以外の、現在の言語学では言語装置を用いて特徴づけすることが難しい側面である。後者の側面を、Halliday ら[5]は首尾一貫性という概念でとらえる。つまり、文章の良さ、自然さに関わる概念は、4.2 節で述べた基本的な言語解析の道具立てでは、不適切な言語表現であることの特徴を合理的に記述できない場合があることを示している。

こういった対象について 4.4 節のような深い言語解析法を導入し、言語分析を多層化・精緻化する方向性もありうるが、本稿は、まず、このような高次の用法添削を 4.2 節の X,Y の記述方法と同時に、Z を不自然と判断した論拠となる記述を提案する。記述例を図 5 に示す。

4.2 節での添削とは違い、図 5 中の添削では、Z は、「【a コーパス】に【(b 指標)】で{類似する／類似しない}ので誤りとする」という形式で、統計情報を利用した判断論拠を記す。つまり、どの資料と比較して、不自然であると判断したかが記述として残る。

日本語の書き言葉だけをとらえても、新聞、雑誌、論文、教科書、随筆、小説など多様である。また、日本語が通時的に変化してきたように、今現在の、日本語も今後変化していくであろう。そのため、どのような表現が一般的に受け入れられる知的な表現であるかは、本質的に主観的な問題であると言える。例えば、自然言語処理では伝統的に、新聞記事をコーパスとして技術を開発してきた伝統があるが、新聞記事が作文添削において「代表性を有する」日本語文章となりうるかは、疑問の残るところである。そこで、あらかじめ、くだけた表現に近い文例を含め複数の文例コーパスを用意しておき、どのサンプルを参考に添削をしたかを記述することを狙う。

複数の類似指標も用意する。図 5 では、bi-gram という語連鎖の指標を例として記述したが、現在自然言語処理では、多様な類似指標が提案されている。例え

ば、Bleu[6]はn語連鎖を元情報として利用して、機械翻訳の良さを評価する。Bleuは自動添削の例ではないが、機械翻訳の分野では、標準的な評価方法として認識されるようになっている。このデータベースの構築により、適切な統計指標も新たに開発できる可能性がある。

文のレベルより、さらに高次の首尾一貫性の誤り指摘も、図5のX,Yの範囲を文内の表現の単位から複数文の単位へ広げることで、拡張できる。文のレベルより高次の首尾一貫性に関連し、かつ、特定のコーパスとの類似度によって評価可能であろう例を挙げておく。

- ていねいさ

です・ます等の丁寧表現の一貫性が保たれているか

- 語彙選択の一貫性

専門用語と一般名詞を混在させていないか、あるいは、漢語、和語、外来語(カタカナ語)の使用が適切か／一貫しているか

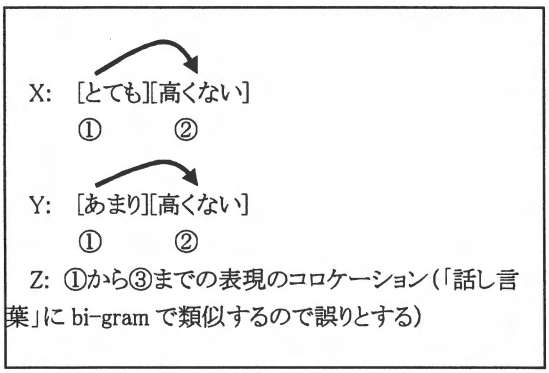

図5. 高次の用例添削の記述例

4.4 誤り／修正事例の特徴の多層化

意味解析の方向性もX,Y,Zの記述の上で利用したい。係り受けは、文節間のグラフを用いて解析結果を表現する、統語解析の一種であるが、同様に、意味解析でも文節間のグラフによって、意味を表現することも採用できる。しかし、意味表現の記述観点は、多様性が高く、本稿では、係り受けと多層表現することが可能な文節間のグラフにより表現する意味表現を挙げるに留める。こういった、文節間を統語的関係のグラフだけでなく、意味的な関係のグラフによって多層に表現することは、添削レコードを作成する際の支援環境に相当の工夫が必要であるが、実現できれば、連体修飾節、省略や照応の問題、「AのB」といった表現の誤りを記述する上で、有力な手がかりとなる。

- 項と項を支配する対象との関係
- 名詞類と名詞類の参照関係

5. まとめ

本稿では、シンポジウムの話題提供として、作文添削事例データベースを協調的に作成する構築案を提案した。

提案した添削の形式的記述を行うためには、図2の②で示した支援環境に関するユーザインタフェースの設計が重要な課題となる。このようなインターフェースは、近年のWeb上電子辞書や翻訳メモリの閲覧方式の工夫が採用できる。また、協調的な作業を行ううえで、blogで採用されているトラックバックのシステムを導入することも有益であろう。

学生の書く作文は、教育の観点からは添削の余地はあるものの、日本語であることは間違いない。Web上やメールで用いられる日本語にはこのような言語運用が溢れている。このようなゆれの存在が、自然言語の本質であり、本稿で提案した添削事例データの蓄積は、柔軟な言語処理技術を開発していく上での基盤データとなることが期待できる。

作文指導は高度な知的活動である以上、人間が指導すべき部分は必ず残るであろう。本稿が提案したようなデータベースの構築は多大な労力が掛かるものであるが、蓄積した添削データを学生の自学自習に役立つシステム構築に役立て、教員が学生とのより本質的な議論・指導に注力できる環境を作り上げていきたい。

参考文献

[1] M. Hearst et al. The Debate on Automated Essay Grading, IEEE Intelligent Systems, v.15 n.5, p.22-37, 2000.

[2] 和泉絵美他.『日本人1200人の英語スピーキングコーパス』,(独)情報通信研究機構,2004.

[3] 大曽美恵子他.日本語学習者の作文コーパス:電子化による共有資源化』研究報告書, 名古屋大学, 1999.

[4] 大野晋.『日本語練習帳』,岩波新書,1999.

[5] M. A. Halliday et al. Cohesion in English London, Longman, 1976.

[6] K. A. Papineni et al. Bleu: a method for automatic evaluation of machine translation. IBM Technical Report, 2001.