

# Comment Transer: 「こめ寅」 ～ 評判分析・機械翻訳技術の Google ガジェット上での実現～

貞光 九月 †      乗松 潤矢 †      福富 崇博 †

† 筑波大学 システム情報工学研究科, {sadamitsu,norimatsu,fukutomi}@mibel.cs.tsukuba.ac.jp

## 1 はじめに

近年、blog をはじめインターネット上に膨大なテキストデータが蓄積されるようになり、それと同時にテキストの中にも含まれる情報を活用することが求められている。我々は現在、他の言語の blog も含めて分析を行う「多言語横断 blog 分析エンジン」の開発を行っており、本稿ではその中間報告として、構成要素となる研究技術を Google ガジェット<sup>\*1</sup>として実装した「Comment Transer:こめ寅」についての概要を述べる。

2 節にシステムの実際の動作についての概説とスナップショットを例示し、3 節以降システムに用いたそれぞれの要素技術について述べていく。

## 2 Comment Transer: 「こめ寅」

本節では我々が作成した Google ガジェット「Comment Transer:こめ寅」(以下「こめ寅」)について述べる。「こめ寅」は基本的には Twitter<sup>\*2</sup>のようにユーザーが日々の雑感(コメント)を記入し、共有するガジェットである。しかし実際に Twitter を見てみると、様々な言語のコメントが混ざり合っており、それら全てを 1 人のユーザーが理解するのは不可能である。また、そのコメントに対する端的な提示がなく、例えばそのユーザーがどのような心境であるのかをアイコンで提示するようであれば、コミュニケーションがより円滑になるのではないかと考えた。そこで「こめ寅」には、ユーザーのネイティブな言語に自動翻訳し、かつ全てのコメントに対して自動的にポジティブ・ニュートラル・ネガティブの感情を付与するという 2 点を大きな特徴として持たせた。図 1 に実際の iGoogle 上での動作画面のスクリーンショットを示す。

ここではユーザーの言語設定をそれぞれ日本語と英語に設定し、2 通り表示している。また Google ガジェットは Google デスクトップや個人の blog にブログパーツとして置くこともできるため、使い次第で様々なことができるのではないかと期待する。

<sup>\*1</sup> <http://www.google.co.jp/ig>

<sup>\*2</sup> <http://twitter.com/>

## 3 Discriminative training を用いた評価文書分類

### 3.1 評価文書分類の概要

ある対象に対する評価を含む文書(評価文書)を、肯定評価・否定評価の 2 値ラベルに分類する評価文書分類 [1] は、その対象に対する評価を定量的に提示できるという点で有益であり広く一般に用いられている。本節では従来の最尤学習 [1] ではなく、discriminative training を導入することで性能の改善を図ると同時に、より精度の良い評価表現辞書を得ることを目的とする。

### 3.2 最小分類誤り学習法

discriminative training のうち代表的なものとして、最小分類誤り (MCE:Minimum Classification Error) 学習法が挙げられる [2]。MCE では数学的に扱いやすいシグモイド関数を損失関数として導入することで、パラメータ集合に関する滑らかな関数を形成できるため、以下の評価関数  $F$  の最小化問題として定式化できるという特徴を持つ。

$$F(\theta) = 1 / [1 + \exp \{ \log p(\omega_c | d; \theta) - \log p(\omega_w | d; \theta) \}]$$

ここで  $\omega_c, \omega_w$  は正解ラベル及び不正解ラベル、 $d$  は文書、 $\theta$  はモデルパラメータである。パラメータの更新には一般最急降下法 (GPD:General Probablistic Descent)[2] を用いて推定を行う。

discriminative training を用いて Amazon<sup>\*3</sup> のレビューデータに対する評価文書分類を行った結果、ベースラインのナイーブベイズ法で 83.90% だった正解率が、88.18% まで大幅に改善した。

## 4 フレーズに基づく統計的機械翻訳

### 4.1 統計的機械翻訳の概要

「こめ寅」の翻訳には、我々の研究室で開発された、階層フレーズに基づく統計的機械翻訳システムを利用している。統計的機械翻訳は、従来の機械翻訳のように人手で翻訳ルールを作成するのではなく、大量の対訳コーパスから自動的にルールを獲得できるため、システムの構

<sup>\*3</sup> <http://www.amazon.co.jp>



図1 「こめ寅」のiGoogle上でのスクリーンショット(仮)

築にかかる人的コストを削減できる。

統計的機械翻訳は基本的には以下の式に従って行われる。

$$\hat{f} = \arg \max_f P(e)P(f|e) \quad (1)$$

ここで  $f, e$  はそれぞれ翻訳元言語の単語列、翻訳先言語の単語列を指し、本システムが、言語モデル ( $P(e)$ )、翻訳モデル ( $P(f|e)$ )、デコーダ ( $\arg \max_f$ ) の3要素から構成されていることを表す。

#### 4.2 階層フレーズモデル

本翻訳システムでは、翻訳の単位を数単語連続したフレーズとし、さらにフレーズを階層的に捉えることでフレーズのペアを CFG(文脈自由文法) の対として表現した Synchronous-CFG を用いている。

前節の翻訳モデルには、翻訳に関する様々な素性が用いられるが、我々は新たな素性としてそれぞれの言語におけるフレーズの出現確率と、フレーズ対の共起確率を加えることにより、BLEU 値 (翻訳精度を表わす指標) において、ベースライン 11.5% に対し、本手法では 12.92% を達成することができた [3]。

### 5 まとめと今後の課題

本稿では我々が開発した評判分析及び統計的機械翻訳技術を用いた Google ガジェット「こめ寅」の機能と、そこに利用されている研究技術について述べた。現段階では翻訳データの著作権の問題上、「こめ寅」をそのま

ま公開することができず、日本語版のみしか公開できないため、今後は著作権の問題をクリアしていくことが1つの課題である。また、パスワード機能を持たせることや、評価文書分類で現在研究を進めているトピック情報の利用についても検討していきたい。

#### 謝辞

本研究の一部は、魅力ある大学院教育イニシアティブ「実践IT力を備えた高度情報学人材育成プログラム」による。

#### 参考文献

- [1] Pang, B. and Lee, L.: Thumbs up? Sentiment Classification using Machine Learning Techniques, *Proc. of the Conference on Empirical Methods in Natural Language processing (EMNLP)*, pp. 76-86 (2002).
- [2] Juang, B.-H. and Katagiri, S.: Discriminative learning for minimum error classification, *IEEE Trans. Signal Processing*, Vol. 40, pp. 3043-3054 (1992).
- [3] 貞光九月, 乗松潤矢, 福富崇博: blogからの自動意見抽出をはじめとする多様なアプリケーションを組み込んだオンラインblog分析エンジンの開発, 筑波大学システム開発型研究プロジェクト2006年度研究成果報告(2007).