統計的機械翻訳における未知の一般語と固有名への対処

水上 仁志　　秋葉 友良
豊橋技術科学大学 情報・知能工学専攻
mizukami@nlp.cs.tut.ac.jp　akiba@cs.tut.ac.jp

1 序論

統計的機械翻訳（SMT）[1] は原言語と目的言語の 2 つの言語への深い知識を必要とすることなく、対訳コーパスから翻訳ルールを自動的に学習する。フレーズベース SMT は、連続する単語列（フレーズ）を翻訳の最小単位として、確率に基づく翻訳候補の順位付けを行い、最も確率の高い候補を出力する。ここで、翻訳に使用するフレーズテーブルが全てのフレーズに対応できれば良いが、学習セットで学習しきれなかった単語は未知語となってしまい翻訳されず、元の言語の文字列のまま出力されてしまうという問題がある。さらに、未知語を含む翻訳には、未知語そのものが翻訳できないだけではなく、未知語を含むようなフレーズに対する翻訳候補がないという問題もある。そのため、未知語によって翻訳が分断されてしまい、周辺のフレーズの翻訳にも影響を与えてしまう。その例を Figure 1 と Figure 2 で示す。Figure 1 では、「developed by」と「Dr.Watson and his colleagues」という 2 つのフレーズが翻訳可能であるのに対して、単語「Watson」が未知語だと仮定した Figure 2 では、「Watson」自体が翻訳されない。それに伴って、全てが既知語である Figure 1 で使用できたアライメントが利用できない。従って、未知語の周辺のフレーズにも誤った翻訳を導くような影響が出てしまう。

未知語を扱った先行研究として、川原らの文字ベースの翻訳手法 [11] があった。この手法は 1 回で翻訳できなかった未知語を文字単位で分割し、それを入力として文字レベルの翻訳モデルによる 2 回目の翻訳を行うという手法であった。しかし、この方法は人名や地名などの固有名に対しては適用できない。固有名の機能は事物を参照することにあるので、必ずしも翻訳先の言語表現に直す必要はなく、翻訳元の文字列のまま翻訳先に現れても許容できる。未知の固有名を扱う手法として、クラスラベルに汎化することで未知語を扱う手法が提案されている。[3][10] 例えば、関らの先行研究 [10] では、フレーズテーブルと言語モデルの地名を全て「PLACE」という文字列へ汎化し、 未知語による翻訳の分断を防いでいた。しかしこれらの研究では、未知の一般語には適用できない。

これらの先行研究を受けて、本研究ではまず、未知語の種類を分類した。そして、分類した未知語の種類に応じて、異なる手法を適用した。固有名の未知語に対しては、固有表現認識器（NER）を利用して、固有名をクラスラベルに置き換える手法を適用し、固有名以外の一般的な未知語に対しては、単語の分散表現による単語間の類似度を用いて、未知語の代わりにそれに類似する代わりの単語を利用した翻訳手法をそれぞれ適用した。また、類似語による翻訳では、入力として単純なテキストの形式の使用に加えて、複数の類似語を翻訳対象として扱える、ラティスの形式も使用して、結果を比較した。

この論文の構成は以下の通りである。まず、2 節で固有表現認識器（NER）を利用した翻訳手法について説明し、3 節で単語間の分散表現による類似語を利用した翻訳手法について説明する。4 節で実験条件と実験結果を示し、5 節で、まとめと今後の課題について述べる。

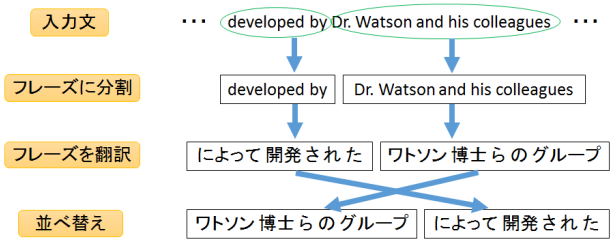

Figure 1: 未知語がないときの一般的なフレーズベース SMT の翻訳の流れ

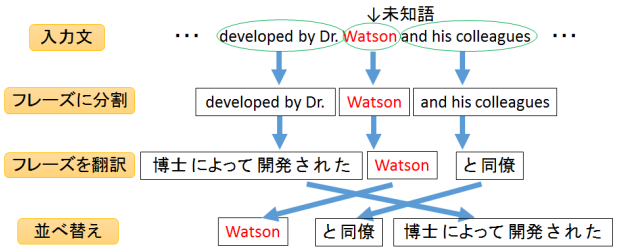

Figure 2: Watson が未知語のときの一般的なフレーズベース SMT の翻訳の流れ

2 固有表現認識器（NER）を利用した提案法

Figure 3 に NER を利用した提案手法の全体の流れを示す。

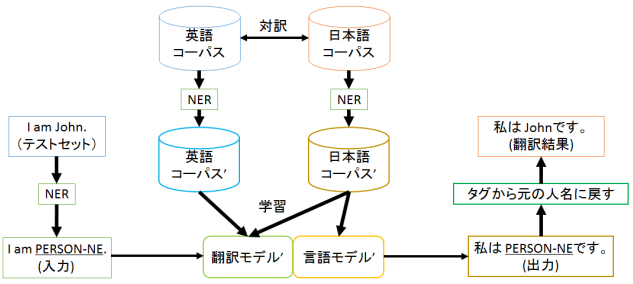

Figure 3: NER を利用した翻訳の流れ

本研究では提案手法を人名の固有名に適用したが、地名や組織名などの他の種類の固有名にも同様に適用可能である。

2.1 学習プロセス

NER を利用した提案法では、学習セットに対して置き換えを行う。人名を例とすると、学習セットに含まれる実際の人名を、「PERSON_NE」のような全て同一の固有名の種類ごとのクラスラベルに置き換えて統一する。置き換えた学習セットによって、モデルを学習する。そのモデルの構築方法を以下で述べる。

1. 学習セットにおける全ての実際の固有名を固有名の種類ごとのクラスラベルに置き換える。そのために、学習セットの原言語側と目的言語側のためにそれぞれの言語に対する NER を使用して、両言語の固有名である単語やフレーズにタグ付けを行う。それによって認識された固有名の位置とその種類の情報によって、両言語側の固有名を全て人名や地名などの種類ごとに単一であるクラスラベルに統一する。人名の場合は、NER でタグ付けされている全ての実際の人名を「PERSON_NE」という同一のクラスラベルに置き換える。クラスラベルに置き換えを行った学習セットの両側を使用して翻訳モデルを、目的言語側のみにより言語モデルをそれぞれ構築する。

2. 学習したフレーズテーブルでは、翻訳前後でクラスラベルの数が異なるフレーズペアも存在してしまう。そのようなペアを翻訳に使用すると、翻訳後の文において固有名、例えば人名の数が増減する文が出現してしまう可能性がある。そのため、それらのペアはフレーズテーブルから削除する。従って、翻訳に使用するのは、フレーズテーブルにおいて翻訳前と翻訳後でクラスラベルの数が同一であるフレーズペアのみである。翻訳に使用するフレーズペアの例を Table 1 に、フレーズテーブルから削除するペアの例を Table 2 に示す。

Table 1: 翻訳に使用するフレーズペアの例

原言語（英語）	目的言語（日本語）
against prime minister PERSON-NE	PERSON-NE 首相 に
against PERSON-NE	PERSON-NE 党首 を

Table 2: フレーズテーブルから削除するペアの例

原言語（英語）	目的言語（日本語）
against prime minister	PERSON-NE 首相 に
against PERSON-NE	党首を

2.2 翻訳プロセス

翻訳プロセスの流れを Figure 4 に示す。

1. 翻訳に用いる入力文に原言語側の NER を適用する。前節の学習プロセス同様に、入力文中の固有名を各クラスラベル（例：「PERSON_NE」）に置き換えた文を SMT の入力として使用する。事前処理によってクラスラベルに置き換えられる前の元の文字列（実際の人名）は以下のステップ 3 で使用するために保持される。

2. 固有名が置き換えられた原言語の入力文を、学習プロセスで構築した言語モデルと翻訳モデルによる統計的機械翻訳のデコーダへそのまま入力する。モデルの学習にも、固有名がそれぞれの種類のクラスラベルに置き換わった学習セットを使用しているため、デコーダの出力も固有名がクラスラベルになっているものとなる。翻訳結果を出力するときに、フレーズアライメントの履歴も生成する。

3. 出力文のクラスラベルは固有名に置き換えを行う前の、元の入力文とフレーズアライメント履歴の情報を使用して、実際の固有名の文字列へ戻す。そのために、デコードプロセスで構築された原言語文と目的言語文のフレーズアライメントの情報を利用する。その情報を用いることで、出力文のクラスラベルから元の原言語文におけるクラスラベルの位置が特定できる。さらに、翻訳プロセスの 1 番目のステップで保持しておいた、元の文字列も使用することで、出力文のクラスラベルを元の固有名に戻す。

3 類似語を利用した提案法

3.1 類似語を利用した翻訳の流れ

固有名以外の一般的な未知語に対して、word2vec[6] による単語の分散表現から得たその未知語に対する類似語を利用した手法を適用する。翻訳の入力にはテキスト形式に加えて、複数類似語を考慮できるラティスの形式も使用する。これにより、未知語が目的言語へ翻訳可能となり、翻訳性能の向上が期待できる。Figure 5 に入力文の「vehicle」が未知語で、「car」が「vehicle」に対する最も類似度の高い単語であるときの置き換え例を示す。

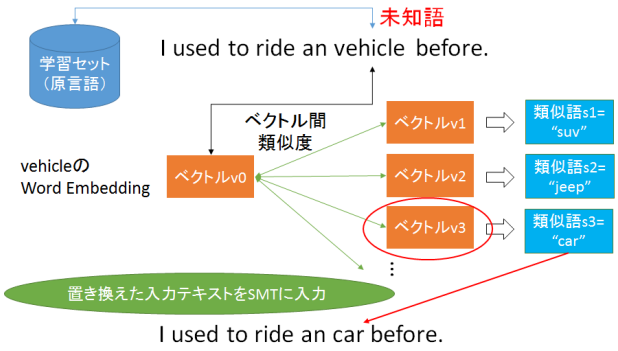

Figure 5: 入力文の未知語を類似語に置き換える流れ

未知語「vehicle」を word2vec の語彙の中で最も類似度が高い「car」へ置き換えて、翻訳を行う。ただし、word2vec の語彙にも含まれない未知語の場合は、類義語を求めることができないので、置き換えは行わない。

3.2 ワードラティス翻訳の利用

前節では未知語に対する類似語を 1 つだけ利用して翻訳を行った。今度は複数の類似語を考慮して翻訳するために、ワードラティスを入力した翻訳を用いる。ワードラティスによる翻訳候補の分岐のイメージを Figure 6 に示す。

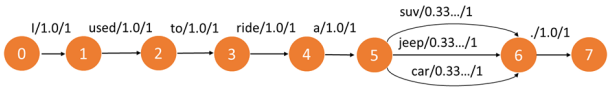

Figure 6: Figure 5 の例による翻訳候補の分岐のイメージ

このラティスを入力に用いることで、様々な翻訳の仮説を考慮した翻訳結果を得る。

4 実験

各手法の有効性を検証するために、ロイターコーパス [9] を使用して実験を行った。全 56782 文のうち、学習セットに 52782 文を開発セットに 2000 文を、テストセット作成用の

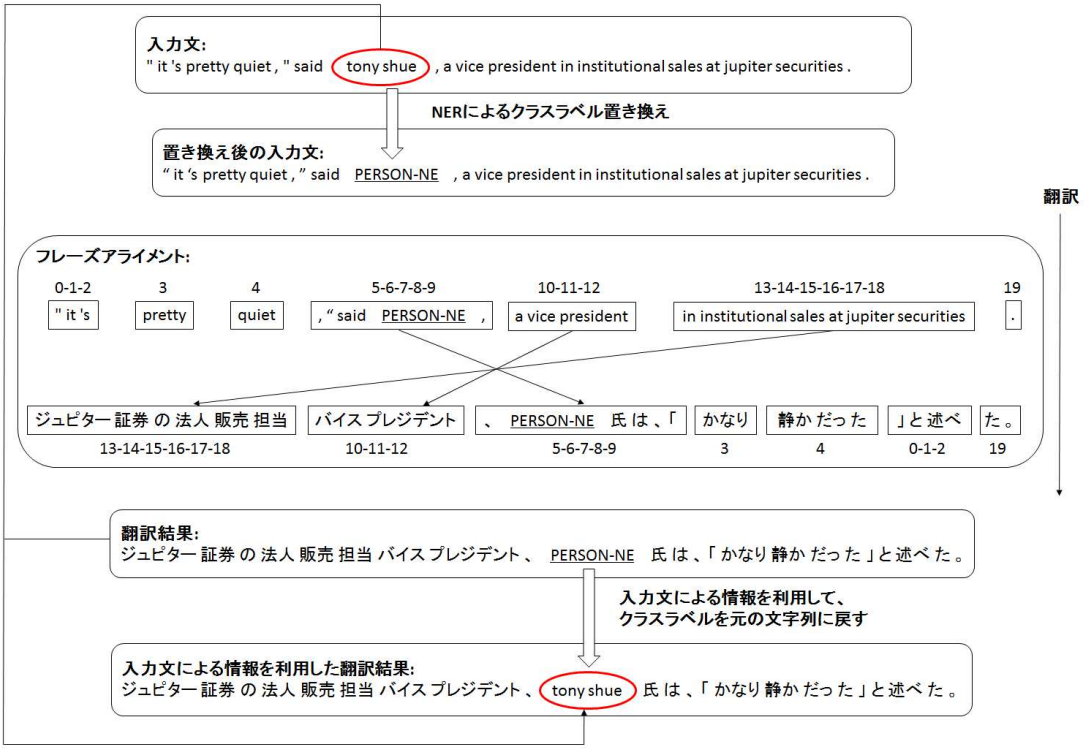

Figure 4: 翻訳プロセスの流れ

データに 2000 文を使用した。実験にはデコーダに Moses[4] を使用し、アライメントツールに GIZA++を使用し、言語モデルは SRILM を用いて学習した。MERT[7] による開発セットを使用したチューニングを行った。学習セットにこれらの手法をそのまま適用して、baseline を構成した。

NER を利用した提案法のために、英語側に Stanford Named Entity Recognizer[2] を、日本語側に CaboCha[5] を使用した。類似語を利用した提案法のための word2vec は GoogleNews データセットの一部を使用して学習された 300 万の単語とフレーズに対する 300 次元のベクトルを含むモデルを使用した。翻訳性能の評価尺度には BLEU[8] を使用した。

4.1 NER を利用した翻訳の結果

提案法の効果を調べるために、テストセットとして用意した 2000 文から、人名を 1 つ以上含む文を抽出して、評価用テストセットとした。その際、人名が学習データに現れないものを 100 文 (以降、未知語文と呼ぶ)、それ以外を 100 文 (以降、既知語文) 選び、合計 200 文を評価した。

提案手法は、固有名を翻訳せず翻訳元言語の表記のままとする。既存手法と対等に比較するために、以下の手順で評価を行った。テストセットの翻訳先文、すなわち正解訳に対し、それに含まれる固有名を翻訳元の固有名表記に置き換えを行い、追加の正解訳を作成した。そして、元の正解役と追加の正解訳を合わせてマルチリファレンスを構成し、これを用いて baseline 及び提案法の評価を行った。以下にマルチリファレンスに用いる正解訳の例を示す。

テストセット

英語側：

with presidential elections set for next week , the white house noted the overall budget deficit had fallen by 63 percent during the four years of the clinton administration .

日本語側（正解訳 1）：

ホワイト ハウス は 、 クリントン 大統領 就任 後 の 4 年 間 で 、 財政 赤字 は 6 3 % の 削減 と なっ た と 発表 し た 。

追加の正解訳

日本語側（正解訳 2）：

ホワイト ハウス は 、 clinton 大統領 就任 後 の 4 年間 で 、 財政 赤字 は 6 3 % の 削減 と なっ た と 発表 し た 。

NER を利用した手法による実験結果を Table3 に示す。

Table 3: NER を利用した翻訳の結果 (BLEU)

	未知語 100 文	既知語 100 文	計 200 文
baseline	18.36	22.69	20.64
提案法	23.05	23.88	23.47
Combine	-	-	23.70

未知語を含む文 100 文で、提案法は従来法を BLEU で 4.69 の改善を得た。また、既知語文 100 文でも効果が見られ、1.19 の改善を得た。これは、未知語ではないものの学習セットでの出現頻度が低い固有名に対しても、提案手法が有効に働いたためと考えられる。既知語文と未知語文を合わせた 200 文において、提案法は 2.83 の改善が得られた。更なる改善を得るために、従来法と提案法を統合し、テスト文に出現する固有名の学習データでの出現頻度に応じて 2 つの手法を切り替えて使用する手法 (Combine) も評価した。具体的には、テストセットの文が含んでいる人名が出現する学習セットの文が 20 文に満たない文は提案法を、そうでないものは従来法を用いた。それによって、合計 200 文で 3.06 の改善が得られた。

4.2 類似語を利用した翻訳の結果

固有名の未知語には NER を利用する手法が効果的であることが分かった。今度は、固有名以外の未知語を対象として、類似語を利用した翻訳実験を行った。次の条件を満たす未知語を含む 100 文をテストセットとして抽出した。

1. 学習セットの原言語側に出現しない
2. 固有名でない。
3. アルファベット以外の文字や記号を含まない。

Table 4: 類似語を利用した翻訳の結果

	baseline	提案法	ラティス翻訳
BLEU	19.05	19.19	19.25

テストセットの未知語を類似度の最も高い 1 単語にテキストを直接置き換えた提案法では、0.14 の BLEU の改善を得た。また、1 つの未知語に対して、類似語の候補として類似度の高い 10 語を選びラティスを作成して翻訳の入力とした時、0.2 の BLEU の改善を得た。

5 まとめと今後の課題

本研究では、未知語を固有名とそれ以外に分類し、それぞれに対する対策法を適用することによって、翻訳性能の改善を達成した。今後の課題として、類似語を利用した翻訳において、単語の分散表現の類似度だけでなく、元の未知語との品詞の一致や、置き換えを行なった後の言語モデル確率を考慮した候補単語選択方法を検討したい。また、固有名ラベルへの汎化手法と類義語への置き換え手法を併用した場合の翻訳性能を調査する予定である。

References

[1] Peter F Brown, Vincent J Della Pietra, Stephen A Della Pietra, and Robert L Mercer. The mathematics of statistical machine translation: Parameter estimation. *Computational linguistics*, Vol. 19, No. 2, pp. 263–311, 1993.

[2] Jenny Rose Finkel, Trond Grenager, and Christopher Manning. Incorporating non-local information into information extraction systems by gibbs sampling. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pp. 363–370. Association for Computational Linguistics, 2005.

[3] Alex Waibel Ian R. Lane. Class-based statistical machine translation for field maintainable speech-to-speech translation. pp. 2362–2365. InterSpeech, 2008.

[4] Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, et al. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th annual meeting of the ACL on interactive poster and demonstration sessions*, pp. 177–180. Association for Computational Linguistics, 2007.

[5] Taku Kudo and Yuji Matsumoto. Japanese dependency analysis using cascaded chunking, proceedings of the 6th conference on natural language learning. *August*, Vol. 31, pp. 1–7, 2002.

[6] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pp. 3111–3119, 2013.

[7] Franz Josef Och. Minimum error rate training in statistical machine translation. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics-Volume 1*, pp. 160–167. Association for Computational Linguistics, 2003.

[8] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pp. 311–318. Association for Computational Linguistics, 2002.

[9] Masao Utiyama and Hitoshi Isahara. Reliable measures for aligning japanese-english news articles and sentences. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics-Volume 1*, pp. 72–79. Association for Computational Linguistics, 2003.

[10] 関拓也, 山本和英. 統計的機械翻訳における地名の汎化の影響. 言語処理学会　第 15 年次大会　発表論文集, pp. 380–383, 2009 年 3 月.

[11] 川原宰, 村上仁一, 徳久雅人. 文字ベース翻訳による未知語処理. 言語処理学会　第 22 年次大会　発表論文集, pp. 207–210, 2016 年 3 月.