意図解釈タスクにおけるスロット表現置換によるデータ拡張

松田 繁樹　　土田 正明　　山上 勝義
株式会社コトバデザイン
{matsuda,tsuchida,yamagami}@cotobadesign.com

1 はじめに

本稿では、対話インターフェースの基本的な機能であるインテントとスロットを認識する意図解釈モデルの為の学習データに対するデータ拡張手法について検討する。特に、学習データが小規模な際にインテント認識に比べて相対的に精度が低くなる傾向にあるスロット認識に着目する。

スロットを高精度に認識するためには、スロットの値となる表現（以降、スロット表現と呼ぶ）の傾向や、スロット表現の周辺文脈の傾向を捉えたモデルの訓練が必要と考えられる。スロット表現自体の傾向を捉えるためには、スロットの種類毎に多様な表現のバリエーションが必要と考えられる。また、スロット認識のための文脈的傾向を捉えるためには、類似した文脈で異なる様々なスロット表現が出現する学習データが必要と考えられる。

本稿では、上記仮説に基づき、1) 多様なスロット表現が含まれ、2) 類似文脈で様々な異なるスロット表現が出現するという２つの要件を満たす学習データを小規模な学習データから自動生成する手法として、学習データ中のスロット表現をスロット表現になりうる類似表現で置換するという単純なデータ拡張方法を検討し、評価した結果を報告する。

2 スロット表現の置換によるデータ拡張

本節では、スロット表現置換によるデータ拡張手法について述べる。テキスト分類タスクのデータ拡張として、単語を一定の割合で類似語で置換する手法 [1] が存在するが、本稿の手法はスロット認識のために単語単位ではなく複数単語からなるスロット表現全体を置換したものと位置付けられる。

まず、学習データ中のスロット表現を抽出し、相互に置換してデータ拡張を行う。次いで、各スロット表現を単語単位に分割し、各単語の類似語の組み合わせもスロット表現候補とし、一定の割合 p_α でその表現候補に置換する。置換するスロット表現候補は以下の２種類の手法で生成する。

Skip-gram[2]　Wikipedia のダンプデータから Skip-gram で分散表現ベクトルを学習し、2 つの単語のコサイン類似度が β 以上の語を類似語とみなす。wikipedia とスロット表現の単語分割は、日本語は Juman++[3]、英語は NLTK の単語分割で行い、各単語の類似語の組み合わせを表現候補とする。

文脈類似語データベース [4]　ALAGIN フォーラムで公開されている、100 万語の名詞に対して、約 1 億ページの Web 文書上での文脈が類似している名詞を類似度順に最大 500 個列挙したデータベース[1] を用いて、見出し語と文字を単語辞書として、動的計画法で分割数最小の基準で単語分割し、見出し語部分を類似度 β 以上の類似語で置換したものを表現候補とする。

また、類似語を用いたスロット表現候補生成によるデータ拡張との比較のために、２つの手法も評価する。

スロット表現の相互置換　学習データからスロット表現の語彙をスロットの種類毎に取得し、種類毎に相互に置換する。すなわち、類似語によるスロット表現候補の生成と置換を行わないデータ拡張である。学習データが小規模な場合、スロット表現の多様性が限定的になるため、スロット表現自体の内容の傾向も周辺文脈の傾向も捉えられないモデルになる可能性がある。

任意の固有名詞による置換　スロット表現の内容にかかわらず、Wikipedia のエントリを固有表現みなして一定の割合 p_α で置換する。任意の固有表現で置換することで、スロットの周辺文脈が認識の手がかりとして強くなると考えられる。

[1] SW_ALAGIN_V1.1.1_1m-rv100k.bbc0.0008.cleaned.data.bz2

ー 文パターンの例 ー

例1. 材料が[材料]で時間が[調理時間]でできる料理
例2. [時間]頃のバス時刻表を表示してね

ー スロット表現の辞書の例 ー

代表語	類義語 1	類義語 2	類義語 3	…
ロゼ	ロゼワイン			
インゲン	隠元豆	菜豆	三度豆	
⋮				

図 1: スロット表現がクラス化された文パターンの例とスロット表現の辞書の例

3 評価実験

3.1 実験コーパス

評価実験は、日本語と英語の 2 つの言語で実施した。日本語は、意図解釈モデルの学習および評価実験を行う為に我々が独自に構築した実験コーパス生成システムを用いて作成した。実験コーパス生成は、28 種類のインテントに対してスロット表現がクラス化された約 8 万文のテキスト (以降、文パターンと呼ぶ) と、54 種類のスロット表現のクラスに対して、約 60 万種類の代表語とその代表語に対する類義語から構成された計約 200 万語彙の辞書を用いて行われる。文パターンとスロット表現の辞書の例を図 1 に示す。カギ括弧で囲まれたスロット表現のクラス文字列を、スロット表現の辞書に含まれる語でランダムに置換することにより、多様なスロット表現を含む実験コーパスを生成することができる。

学習データと評価データはお互いにオープン条件になるように、文パターンと、スロット表現の代表語（及びその代表語の類義語）を選択し、前述の実験コーパス生成システムを用いて学習用及び、評価用データを生成した。学習データは、約 3 万 6 千種類の文パターンと、語彙サイズ約 3 万のスロット表現の辞書から 65536 文を生成した。小規模な学習データしか利用できない条件での評価では、これらの中から選択した 256, 512, 1024, 2048 文章を使用した。十分な量の学習データが利用可能な場合の評価として、32768, 65536 文を用いた場合の実験を行った。開発データは、学習データと同じ文パターンとスロット表現の辞書を用いて 256 文だけ生成した。開発データ量は、学習データ量と同様に小規模である。評価データは、学習データに含まれない文パターンと代表語 (及びその代表語の類義語) を用いて 5000 文を生成した。

英語の実験コーパスとして、パーソナルボイスアシスタントで収集された SNIPS[5] を使用した。SNIPS は、7 種類のインテントと 39 種類のスロットを含む 700 文の評価データ、700 文章の開発データ、13084 文の学習データから構成されている。日本語と同様に、小規模な学習データしか利用できない条件での実験を行うため、64, 128, 256, 512 文を学習データとして使用した。十分な量の学習データが利用可能な場合の評価として、8192, 13084 文を用いた場合の実験を行った。開発データはランダムに選択した 64 文である。評価データは 700 文である。

3.2 認識モデル

インテントとスロットを認識する意図解釈のタスクにおいて高い認識性能が報告されている BERT ベースのモデル [6] を用いて評価実験を行った。単語への分割及び、必要に応じて BPE (Byte Pair Endoding) により分割されたサブワード列が BERT に入力され、インテントの種類に続き、個々のサブワードに対するスロットの種類が出力される。

日本語の実験では、BERT の事前学習モデルとして、京都大学の黒橋･河原研究室が公開している BERT 日本語 Pretrained モデル[2] を使用した。日本語の Wikipedia のテキストを用いて学習されている。形態素解析には Juman++ を用いた。SNIPS を用いた英語の実験では、BooksCorpus と英語 Wikipedia を用いて学習されたモデル[3] を事前学習モデルとして用いた。これらのモデルは、12 層、各層のユニット数 768, 12 個のマルチヘッドアテンションの構造を持つ。

3.3 実験結果

3.3.1 データ拡張無し

日本語の実験結果を図 2 の上表、SNIPS（英語）の実験結果を同図の下表に示す。評価尺度は macro F 値を用いた。データ拡張を全く行わない場合の評価結果を「データ拡張無し」の列に示す。

日本語において、学習文章数が 256 の場合、インテント当たりの文パターン数は平均 9.1 (= 256/28) 、スロット表現の種類毎の平均語彙数は平均 7.3 (= 392/54) であり非常に少ない。認識性能が飽和したと考えられる学習文章数 32,768 の時のスロットに対する F 値が

[2] http://nlp.ist.i.kyoto-u.ac.jp
[3] https://github.com/google-research/bert

図 2: 日本語と英語の実験結果

日本語の実験結果

学習文章数	文パターン数	スロット語彙サイズ	データ拡張無し		スロット表現の相互置換		任意の固有名詞による置換			Skip-gram				文脈類似語データベース			
			macro F 値		macro F 値		h.p.	macro F 値		h.p.		macro F 値		h.p.		macro F 値	
			intent	slot	intent	slot	p_α	intent	slot	p_α	β	intent	slot	p_α	β	intent	slot
256	254	392	96.26	37.44	96.54	42.66	0.2	96.77	63.25	0.6	0.5	95.90	64.32	0.4	0.1	97.62	65.19
512	500	787	98.08	58.83	98.82	66.18	0.2	98.40	74.30	0.8	0.6	97.98	78.16	0.2	0.2	98.72	78.22
1,024	977	1576	99.00	78.60	99.00	82.77	0.2	99.18	86.68	0.7	0.7	99.00	88.78	0.2	0.0	99.20	86.43
2,048	1,899	3149	99.38	88.33	99.34	91.85	0.2	99.68	91.56	0.7	0.7	99.60	93.02	0.4	0.3	99.70	93.37
32,768	21,510	22171	99.84	96.76													
65,536	37,225	24201	99.88	97.10													

英語 (SNIPS) の実験結果

学習文章数	文パターン数	スロット語彙サイズ	データ拡張無し		スロット表現の相互置換		任意の固有名詞による置換			Skip-gram			
			macro F 値		macro F 値		h.p.	macro F 値		h.p.		macro F 値	
			intent	slot	intent	slot	p_α	intent	slot	p_α	β	intent	slot
64	64	130	88.14	35.50	86.86	38.11	0.2	81.71	37.01	0.2	0.3	89.29	38.12
128	124	246	95.43	49.29	94.14	51.64	0.2	92.86	53.45	0.2	0.6	95.57	60.40
256	243	428	97.00	62.41	96.71	64.25	0.2	93.71	61.02	0.2	0.5	97.00	67.51
512	461	747	98.29	75.89	96.14	75.89	0.2	96.71	72.18	0.2	0.5	97.57	77.82
8,192	5,439	7761	98.86	91.68									
13,084	7,976	11514	99.14	91.04									

96.76 なのに対して学習文章数 256 では 37.44 と大幅に劣化していることがわかる。一方、インテントに対する F 値はスロットに対する F 値に比べて劣化幅が小さいことがわかる。英語においても同様の傾向が見られる。

3.3.2 スロット表現の相互置換

学習データに含まれていたスロット表現をスロットの種類毎に取得し、種類毎に相互置換を行った場合の評価結果を表中の「スロット表現の相互置換」の列に示す。日本語の評価結果において、データ拡張無しと比較したスロット表現の相互置換によるデータ拡張の効果は、学習文章数 256, 512, 1024, 2048 の各々について、約 8.3%, 17.8%, 19.5%, 30.2% と文章数が増加するに従って F 値の上昇率が大きくなっていることがわかる。(F 値の上昇率は $8.3 = (42.66 - 37.44)/(100 - 37.44)$ で算出) 英語の評価結果におけるスロット表現の相互置換によるデータ拡張の効果は、学習文章数 64, 128, 256, 512 の各々について、約 4.0%, 4.6%, 4.8%, 0.0% となり、学習文章数 512 の時にデータ拡張無しと比較して F 値の改善が得られなかった。

これらの実験結果から、学習文章数が少ない場合や、十分な量の学習データが利用可能な場合にはスロット表現の相互置換によるデータ拡張の効果は小さくなり、一方、中間的な学習データ量においてはデータ拡張の効果が得られる傾向が見られることがわかる。

3.3.3 任意の固有名詞による置換

スロット表現の種類に関わらず、Wikipedia のエントリを固有表現とみなして一定の割合でスロット表現を置換することによるデータ拡張の評価結果を、表中の「任意の固有名詞による置換」の列に示す。前述のスロット表現の相互置換によるデータ拡張を行った学習データに対して本手法を適用した。従って比較対象となるベースライン手法は、スロット表現の相互置換である。後に述べる Skip-gram や文脈類似語データベースを用いた実験も同様である。日本語の評価結果の表に示すように、スロット表現の相互置換と比較して、学習文章数が少ない場合に F 値の上昇が見られることがわかる。一方、英語の評価結果においては、全般的にスロット表現の相互置換と比較して F 値が低下していることがわかる。本実験で使用した日本語の実験コーパスは助詞などの比較的明示的な文脈の手がかりがあるため、前後の文脈からスロット表現の種類を推測しやすかったことが考えらえる。それに対して、英語の実験で使用した SNIPS は、スロット表現が連続するテキストが多く含まれており (たとえば、”play <artist><music_item>tunes” など) 文脈のみからス

ロットの種類を推測することが比較的困難であったことが考えられる。

これらの実験結果から、文脈によるスロットの種類の推測が難しいと考えらえるデータに対しては、任意の固有名詞で置き換えるのではなく、スロット表現の種類毎の傾向を崩さないように類似語等により置換するデータ拡張が必要であることがわかる。

3.3.4 Skip-gram で抽出した類似語による置換

Skip-gram により得られた類似語を用いて置換することによるデータ拡張の効果を評価した。日本語の Skip-gram の学習データは、Wikipedia のテキストに対して Juman++ により形態素解析を行い単語分割を行った。また、英語の Wikipedia のテキストに対しては、 NLTK により単語分割を行った。頻度 4 以下の単語についてカットオフを行い、320 次元の分散表現ベクトルを学習した。日本語の語彙サイズは約 64 万、英語の語彙サイズは約 78 万である。

実験結果を、表中の「Skip-gram」に示す。表に示すように、日本語、英語の両方において、スロット表現の相互置換及び、任意の固有名詞による置換と比べて全ての場合で、スロットに対する F 値が上昇していることがわかる。特に英語において、任意の固有名詞による置換では F 値の上昇は得られなかったのに対して安定的に上昇していることがわかる。

これらの実験結果から、スロットの種類毎の表現の傾向が保持されると考えられる類似語による置換によるスロット表現のデータ拡張は、スロットに対する F 値の上昇にとって効果があることがわかる。

3.3.5 文脈類似語データベースの使用

最後に、文脈類似語データベースを用いた類似語によるデータ拡張の効果を評価した。表中の「文脈類似語データベース」の列に評価結果を示す。文脈類似語データベースは日本語の単語に対するデータであるため、実験は日本語のみである。表に示すように、任意の固有名詞による置換と比較して全般的に高い F 値が得られていることがわかる。

文脈類似語データベースは、Skip-gram による方法とは異なり、係り受け関係からの類似語の推定や、人手による整備などが行われており類似語の推定の精度が高いことが期待されたが、Skip-gram と同程度のデータ拡張の効果であった。

4 まとめ

本稿では、意図解釈モデルのためのスロット表現に対するデータ拡張の効果について、スロット表現の相互置換、任意の固有名詞による置換、Skip-gram により得られた類似語による置換、文脈類似語データベースの類似語による置換の、4 種類のデータ拡張法の評価を行った。評価結果から、小規模な学習データしか利用できない場合において、類似表現で置換することによるデータ拡張によりスロットに対する F 値が上昇することを確認した。

Skip-gram と異なり、文脈類似語データベースの類似語リストは、係り受け関係などのより高次の情報を用いて作成されている。従って、これら 2 つの手法を組み合わせることにより、お互いのデータ拡張の効果を最大限に活用することができる可能性がある。今後は、このような 2 つの手法の組み合わせの評価や、スロット表現だけでなく文脈自体のデータ拡張について評価を行う予定である。

謝辞 本稿の一部は、平成 31 年度総務省委託研究「高度対話エージェント技術の研究開発・実証」の成果によるものである。

参考文献

[1] Jason W Wei and Kai Zou. Eda: Easy data augmentation techniques for boosting performance on text classification tasks. *arXiv preprint arXiv:1901.11196*, 2019.

[2] 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.

[3] Hajime Morita, Daisuke Kawahara, and Sadao Kurohashi. Morphological analysis for unsegmented languages using recurrent neural network language model. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pp. 2292–2297, 2015.

[4] Jun'ichi Kazama, Stijn De Saeger, Kow Kuroda, Masaki Murata, and Kentaro Torisawa. A bayesian method for robust estimation of distributional similarities. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pp. 247–256. Association for Computational Linguistics, 2010.

[5] Alice Coucke, Alaa Saade, Adrien Ball, Théodore Bluche, Alexandre Caulier, David Leroy, Clément Doumouro, Thibault Gisselbrecht, Francesco Caltagirone, Thibaut Lavril, et al. Snips voice platform: an embedded spoken language understanding system for private-by-design voice interfaces. *arXiv preprint arXiv:1805.10190*, 2018.

[6] Qian Chen, Zhu Zhuo, and Wen Wang. Bert for joint intent classification and slot filling. *arXiv preprint arXiv:1902.10909*, 2019.