QA レポート

システム情報工学研究科 2 年　200820634　越川 満

研究題目：統計的機械翻訳におけるフレーズ対応最適化を用いた翻訳候補のリランキング
指導教員：山本 幹雄
発表日時：2009 年 5 月 14 日

● 評価指標 BLEU についての質問

【 質問 】
（1）BLEU とはどのようにして求められるのか
（2）またその数値は何を意味しているのか（例えば BLEU 30%とは？）
[発表時の回答]
（1） BLEU は翻訳結果と正解翻訳例との一致率を測る指標であり、人間の感覚と高い相関を示している。BLEU は具体的には、翻訳結果と正解翻訳例の一致の割合を、1 単語ずつ比較した場合、連続する 2 単語ずつ比較した場合、･･･、連続する 4 単語ずつ比較した場合について求め、その幾何平均を取る評価指標である。
（2） BLEU 値 30%とは、翻訳システムが 3 割の確率で正解翻訳例と同じ結果を出力するという意味ではなく、翻訳システムが出力した結果の単語列と正解翻訳例の単語列が 3 割程度一致するということを表す。
[改善した回答]
（1） （発表時の回答に加えて）より詳細な BLEU の計算方法については、[1]を参照。
（2） 発表時の回答と同じ

【 質問 】
提案手法による BLEU 改善は 0.2%であるが、この改善は人間が実感できるのか。
[発表時の回答]
0.2%の改善は翻訳結果を見てもほとんど実感することができない。
BLEU の改善が 1～2%になって初めて翻訳精度の改善が実感できる。
[改善した回答]
これまでの経験から、0.2%の改善は翻訳結果を見てもほとんど実感することができず、BLEU の改善が 1～2%になって初めて翻訳精度の改善を実感することができるといった印象である。

【 質問 】

提案手法では、どの程度の性能改善が見込めたのか。

[発表時の回答]

性能改善の上限は BLEU で 5%程度見込めることが知られている。

[改善した回答]

翻訳システムにより選ばれた翻訳候補上位 N 個の中には、システムの出力となる確率最大の候補よりも高い BLEU を示す候補が含まれていることが知られている[2]。テストセットの各文に対して、翻訳候補上位 N 個の中から最も高い BLEU を示す候補を選ぶと、全体として 5%ほど BLEU が向上されるという結果が[2]により報告されている。

【 質問 】

実験結果のグラフで、ビーム幅に対して BLEU が単調増加している様子が見られるが、こういった傾向は一般に見られるのか。

[発表時の回答]

探索精度：ビーム幅・確率と翻訳精度：BLEU の間には強い相関が見られることが知られており、今回の結果は妥当であると考えられる。

[改善にした回答]

（発表時の回答に加えて）なお、システムが選んだ翻訳候補の確率と BLEU の相関については[3]に詳細が書かれている。

- 学習データ（コーパス）についての質問

【 質問 】

コーパスとは具体的にどのようなものなのですか

[回答]

コーパスとは、自然言語（人間が話したり、読み書きしたりする言語）を大量に集めたテキストデータのこと。

【 質問 】

（Web 上のテキストで学習する場合）Web サイト上のテキストには文法を間違ったものや辞書にないスラング等が含まれているが、これは学習に際して問題ないのか

[回答]

実際に Web 上のデータを学習データとして利用する場合、文法間違いやスラングは確率的翻訳規則の学習時に ”ノイズ” として悪影響を及ぼすと考えられる。しかし、文法間違いやスラングを含む文を除く前処理を施すことで、Web のテキストを学習データとして利用可能になると思われる。なお対訳を作ることが難しいため、Web 上のテキストを利用した学習データに作成ついての研究は、現在のところあまり盛んではない。

- 提案手法についての質問

【 質問 】

計算時間はどのように求めるか。また、計算時間を改善できるか（提案手法は実用化できるか)

[発表時の回答]

提案手法は、まずベースラインシステムにより翻訳を行い、その結果得られた翻訳候補上位 N 個をフレーズ対応最適化しリランキングする。すなわち提案手法の計算時間は、ベースラインシステムによる翻訳時間とフレーズ対応最適化時間の和となります。また提案手法による翻訳精度改善は非常に小さく、同じ翻訳精度を得るときにはベースラインシステムの方が翻訳速度は速くなります。したがって、実用上ベースラインシステムで十分です。

[改善した回答]

発表時の回答と同じ

【 質問 】

同じ意味であるが言い換えされている場合はどのように考慮されるのか

[回答]

統計的機械翻訳では、翻訳システムは意味レベルまで踏み込んだ解釈は行わない。したがって、同じ意味であるが言い換えがされている場合も異なる翻訳結果であると判断される。

【 質問 】

ビーム幅が大きくなるとベースラインと提案手法の翻訳精度 BLEU がほぼ同じになる理由はなぜか

[回答]

ビーム幅は翻訳時の探索範囲の大きさを決めるパラメータであり、ビーム幅が大きいほどより多くの翻訳候補（生成される目的言語文とそのフレーズ対応の組合せ）を探索する。ビーム幅が 500～1,000 と大きいときにベースラインと提案手法とで BLEU がほぼ等しくなるのは、ベースラインが翻訳時に各翻訳候補の目的言語文に対して付与したフレーズ対応がその翻訳候補に対する最適なフレーズ対応であり、提案手法によるフレーズ対応改善の余地がない場合がほとんどであったためである。

- 反省点

【 良かった点 】

- 話す速度は遅めだったが、研究テーマが難しいので、発表内容を理解するのにちょうどよい速度であったと思われる。
- 発表中に聴衆を見て話せていたので、この点はこれからも心がけていきたい。
- ややオーバーしたが設定された発表時間で発表できた

【 改善が必要な点 】

- 指し棒を使いすぎており、聴衆にやや煩わしい印象を与えたかもしれない。今後は、指し棒で指す内容を重要なもののみとして、注目すべき点がどこなのかはっきりとさせたい。
- 図が込み入っており、わかりづらいものがあった。また抽象的な説明の図が多く、その反面具体例が少ないため、異分野の研究者には理解しづらかったと思われる。
- 発表時間のうち、評価実験結果に対して割いた時間が少なく、評価指標および結果についての考察が伝わりにくかった。
- 教室の広さを考えるとやや声が小さかったと思われる。

- 参考文献

[1] K.Papineni, S.Roukos, T.Ward and W.Zhu. 2001. “BLEU: a method for automatic evaluation of machine translation.”, In *ACL '02: Proceedings of the 40th Anuual Meeting on Association for Computational Linguistics*, pp.311－318.

[2] S.Hasan, R.Zens and H.Ney. 2007. “Are Very Large N-Best Lists Useful for SMT?”, In *Proceedings of Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Companion Volume, Short Papers*, pp.57－60.

[3] R. Zens and H. Ney. 2008. “Improvements in Dynamic Programming Beam Search for Phrase-based Statistical Machine Translation.”, In *Proceedings of International Workshop on Spoken Language Translation*, pp.195－205.