（ 続紙 1 ）

京都大学	博士（情報学）	氏名	Prasanna Raj Noel Dabre
論文題目	Exploiting Multilingualism and Transfer Learning for Low Resource Machine Translation（低リソース機械翻訳における多言語性と転移学習の活用）		

（論文内容の要旨）

This thesis is about various techniques to improve machine translation (MT) for low resource languages, exploiting multilingualism and transfer learning. The focus was on methods that improve the performance of statistical as well as neural MT systems. This thesis quantitatively shows how using multiple related languages and transfer learning is helpful in pivot language MT, multi-source MT, and domain adaptation. The thesis consists of 7 chapters as described below.

Chapter 1 introduces various concepts and MT paradigms which the thesis revolves around. The chapter starts with a brief history of the evolution of MT, right from rule based to statistical to neural MT. At the end of the chapter is an overview of the thesis and its contributions to the field of MT. The importance and the reason behind the title of the thesis and its contents is explained at the end of the chapter.

Chapter 2 starts with a preliminary investigation of using multiple languages in pivot language MT for Japanese-Hindi. Techniques like phrase table interpolation and multiple decoding paths have been explored in order to utilize multiple translation models obtained using different pivot languages. Languages that are linguistically similar to the source or target languages have been shown to help improve translation quality by around 2 BLEU points. It is also seen that using multiple languages leads to additional improvements in BLEU.

Chapter 3 shows how using noise control methods and neural networks can help obtain large, high quality, technical domain Chinese-Japanese dictionaries. Noise control methods are shown to be useful because pivot language approaches lead to noisy translation models with many inaccurate translation examples. Statistical significance pruning helps improve the recall of the dictionaries by around 2% and makes translation tables up to five times smaller. This makes them practically useable during deployment due to reduced memory requirements and faster lookup. Neural network approaches help in obtaining useful features for re-ranking the outputs of the MT systems leading to higher recall. The outcome is a large Chinese-Japanese dictionary of 3.6 million entries which will be released. Manual investigations indicate that 90% of this dictionary contains entries of perfect quality.

Motivated by the success of neural approaches, Chapter 4 explores ways to develop domain specific neural machine translation (NMT) models of high quality. Despite the limitations of NMT in resource scarce scenarios, this chapter focuses on methods that can leverage external domain data and monolingual data to improve the NMT performance. Techniques like “mixed fine tuning” which jointly learn a multi domain model are proposed and show an improvement of around 10 BLEU points for the resource scarce domain. This approach focuses on joint learning and fine tuning as a way to leverage already learned parameters and hence speed up the NMT training process. Monolingual data is incorporated into the approaches by either shallow fusion or by translating them into synthetic corpora.

Chapter 5 expands on the transfer learning approaches to include multilingualism. It is observed that using multiple languages from the same language family significantly boosts the translation quality for low resource language pairs. Exhaustive experimentation shows that using a related language is always better than using an unrelated one. The chapter focuses on translation to and from English for over six low resource languages. Self-learning approaches focusing on synthetic data were also studied. A particular scenario where a translation system with English as the target language generates synthetic data and incorporates this data to further improve translation to English shows promise. Visualizations of the continuous space representations generated by the multilingual NMT models show that joint learning leads to similar representations for the same sentence across languages. This indicates that the resource poor language can leverage the representations of the resource rich language leading to improved translations. This chapter also shows that it is important to have multilingual corpora because they can act as backups to improve translation quality in case large domain specific bilingual corpora are not available.

Chapter 6 observes the power of black-box NMT approaches for multi-source NMT. Concatenating the same sentence in multiple languages gives improvements of up to 5 BLEU points for resource poor Indian languages. For resource rich European languages improvements of up to 3 BLEU points are observed. Multilingual models are shown to improve performance of bilingual models by up to 2 BLEU points by initializing the bilingual model parameters using the multilingual model parameters. This is related to the concept of knowledge distillation. The working of the NMT model in such a scenario is studied and a method was proposed to extract a multilingual dictionary using the attention mechanism. The crucial observation is that an NMT model is clever enough to identify sentence boundaries and synonyms automatically.

Chapter 7 concludes the thesis. This thesis emphasizes that multilingual multi-way corpora where the same sentence is available in multiple languages should be the focus of language corpora development. As for the future work this thesis discusses the latest NMT architectures and how they will work with transfer learning approaches.

注)論文内容の要旨と論文審査の結果の要旨は１頁を３８字×３６行で作成し，合わせて，３，０００字を標準とすること．
論文内容の要旨を英語で記入する場合は，４００～１，１００words で作成し
審査結果の要旨は日本語５００～２，０００字程度で作成すること．

（続紙 ２ ）

（論文審査の結果の要旨）

本論文は、大規模な対訳コーパスが存在しない低リソースの状況において機械翻訳を改善する方法を提案するものである。具体的には多言語性と転移学習を活用する方策を提案しており、得られた主要な成果は以下の通りである。

１．科学技術分野の日中対訳辞書を英語を介して構築する上で、統計的有意性に基づく日英、英中のフレーズテーブルの枝刈りと、ニューラルネットワークに基づくリランキングによって辞書構築の再現率が向上することを示した。この結果、人手評価で精度約 90%、360 万エントリーの日中対訳辞書の構築に成功した。

２．特定のドメインにおいて対訳コーパスが十分でない場合、ニューラル翻訳の精度は非常に低下する。この問題を解決するために、別ドメインの大規模対訳コーパスを用いてニューラル翻訳のパラメータを初期化すること、および、両ドメインのコーパスを混合して単一のニューラル翻訳モデルを学習することにより、コーパスの少ない特定ドメインの翻訳の BLEU 値を約 10 ポイント改善できることを示した。さらに、目的言語の単言語コーパスからニューラル翻訳によって原言語文を得て疑似対訳コーパスとして利用することによりさらに翻訳精度が向上することを示した。

３．低リソースの翻訳において、言語横断的な転移学習が可能であること、特に類似言語の対訳コーパスの利用が有効であることを、英語と他の 6 言語間の翻訳の大規模な実験において示した。また、多言語のニューラル翻訳の内部のベクトル表現を可視化して分析することにより、同一の文に対して言語の違いを越えた近い表現が獲得されていることを明らかにし、低リソース言語ペアの翻訳において高リソース言語ペアの内部表現が活用されていることが翻訳精度向上に寄与していることを確認した。

４．低リソースの状況においても高リソースの状況においても、多言語の同一内容の文を単純に連結して入力文とする形でニューラル翻訳の訓練を行うことにより、翻訳精度が大幅に向上することを示した。また、通常の二言語間の翻訳においても、このように多言語入力で学習したモデルのパラメータを初期値とすることで翻訳精度が向上することを示した。さらに、翻訳システムの注意機構の内部状態から多言語辞書の構築が可能であることを示した。

よって、本論文は博士（情報学）の学位論文として価値あるものと認める。また、平成３０年２月１５日に実施した論文内容とそれに関連した試問の結果合格と認めた。

注)論文審査の結果の要旨の結句には、学位論文の審査についての認定を明記すること。
更に、試問の結果の要旨（例えば「平成　年　月　日論文内容とそれに関連した口頭試問を行った結果合格と認めた。」）を付け加えること。

Ｗｅｂでの即日公開を希望しない場合は、以下に公開可能とする日付を記入すること。

要旨公開可能日：　　年　　月　　日以降