（ 続紙 1 ）

京都大学	博士（情報学）	氏名	須藤　克仁
論文題目	A Japanese-to-English Statistical Machine Translation System for Technical Documents （技術文書に対する日英統計的機械翻訳システム）		

（論文内容の要旨）

This thesis addresses a Japanese-to-English statistical machine translation (SMT) system for technical documents. Machine translation (MT) is a promising solution for growing translation needs. Japanese-to-English MT is one of the most difficult language pairs due to their large lexical and syntactic differences. This thesis work focuses on patents as the most demanded technical documents that have different attributes from other general documents: technical terms and long complex sentences. This thesis tackles three important research problems in the target task: word segmentation on technical terms, unknown katakana word transliteration, and long-distance reordering. Novel techniques are proposed to overcome these problems: domain adaptation of word segmentation using very large-scale patent data, noise-aware translation fragment extraction for accurate machine transliteration, and syntax-based post-ordering for efficient and accurate long-distance reordering.

Chapter 2 gives a brief introduction of SMT techniques on which the proposed methods are based. They are established and widely used in various language pairs but not sufficient for the Japanese-to-English patent SMT.

Chapter 3 presents a novel domain adaptation method for the Japanese word segmentation, using very large-scale Japanese monolingual unlabeled corpora. The proposed method utilizes word boundary clues called Branching Entropy and pseudo-dictionary features obtained from the Japanese monolingual corpora. The probabilistic characteristic of the Branching Entropy mitigates the stability issue of the baseline method using Accessor Variety. The method achieved word segmentation F-measure of 98.36% and out-of-vocabulary word recall of 92.61% in word segmentation experiments, which were significantly higher than the performance of the baseline methods.

Chapter 4 presents a novel noise-aware character alignment method which extracts meaningful transliteration fragments. Although more than a half of unknown words in Japanese-to-English patent SMT are katakana words, they can be translated into the original English words. However, the transliteration is not straightforward because of the ambiguous and inconsistent mapping between katakana and English phonemes. This work focuses on partial noise in transliteration candidates extracted from the bilingual corpora to learn the mapping, which has not been addressed by previous studies that model sample-wise noise only. The proposed method achieved transliteration accuracy of 66% for unknown katakana words, which is 10% error reduction from the method addressing sample-wise noise only.

Chapter 5 presents a novel efficient SMT method called post-ordering that divides the SMT problem explicitly into two steps: monotone lexical translation by the phrase-based SMT and reordering by the syntax-based SMT. The post-ordering approximates the accurate but computationally expensive syntax-based SMT by using an intermediate language with English words in the Japanese word order. The post-ordering achieved accurate translation comparable to the syntax-based SMT with more than six-time faster decoding speed.

Chapter 6 presents a patent SMT system integrating the techniques presented above. This system has two major advantages other than the advantages of the individual techniques. First, domain adaptation of Japanese pre-processing is needed only on word segmentation, not on more difficult Japanese parsing. Second, katakana unknown words are translated prior to reordering and expected to be reordered correctly without special treatment of unknown word reordering. The system achieved the BLEU scores of 34.77% and 35.75% for the NTCIR-9 and NTCIR-10 PatentMT test sets, which were consistently higher than the performance of the baseline systems using the standard techniques.

Chapter 7 concludes the thesis. The proposed SMT framework realizes a practical Japanese-to-English SMT system adapted to technical documents, where many technical terms and long sentences cause serious translation errors. The propsoed methods do not rely on additional human annotations on in-domain corpora, and can be trained with existing bilingual and monolingual corpora. Finally, some further prospects are discussed.

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、特許等の技術文書を対象として日本語から英語に自動翻訳するシステムに関する研究をまとめたものである。日英の機械翻訳は、語彙や文法の違いが大きいことから今でも困難であるが、技術文書では専門用語が多く、文も長いという問題が加わる。本研究ではこれらに対処するために、専門用語の単語分割、カタカナの未知語の翻字（英文字列への変換）、そして長距離のフレーズの並び換えの問題に取り組んでいる。具体的に得られた主な成果は以下の通りである。

1. 日本語の単語分割において、大規模な教師ラベルなしデータを用いた学習法を提案した。具体的には、教師なしで得られる分岐エントロピと擬似辞書の素性を用いることで、従来手法より高い単語分割精度(98.36%)及び未知語の検出率(92.61%)を実現した。

2. カタカナ語の英語への翻字において、曖昧性やノイズに頑健な手法を提案した。対訳コーパスから自動抽出されたフレーズ対において、部分的なノイズを考慮しながら、カタカナ文字列と英文字列の対応付けをベイズ学習する方法を定式化・実装し、従来手法に比べて、誤り率を10%以上削減し、翻字精度66%を実現した。

3. 日英の統計的機械翻訳(SMT)において、まずフレーズ単位の翻訳を行った上で、統語情報に基づいて事後並び換えを行う方式を提案した。この方式は、日本語と英語の構文的特性を考慮したもので、従来の統語情報に基づくSMTと比べて、6倍以上の処理速度で同等の翻訳精度を実現した。

以上の手法を統合して、日英特許翻訳システムの実装・評価も行っており、本論文は、学術上・実用上寄与するところが少なくない。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26年12月26日に論文とそれに関連した内容に関する口頭試問を行った結果、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