TYPE OF INDUSTRY

科学技術・大学

情報通信研究機構

NICT 先端研究 ⑧⑧

アブギダ系文字 入力効率化

先進的音声翻訳研究開発推進センター・先進的翻訳技術研究室テニュアトラック研究員 丁 塵辰（てい じんしん）

15年3月筑波大学大学院修了。同年4月よりNICTに勤務、有期研究員。18年4月よりテニュアトラック研究員、現在に至る。自動翻訳、アジア言語の整備・解析に関する研究に従事。博士（工学）。

NICTでは自動翻訳への活用などのため、アジア言語における文字・音声処理技術に関する研究開発を行っている。私はその中でもアブギダ系文字を効率的な入力を実現するため、計算言語学理論と言語処理技術を併用し、文字体系にある冗長性を低減して、使いやすい入力インターフェースを創出した。

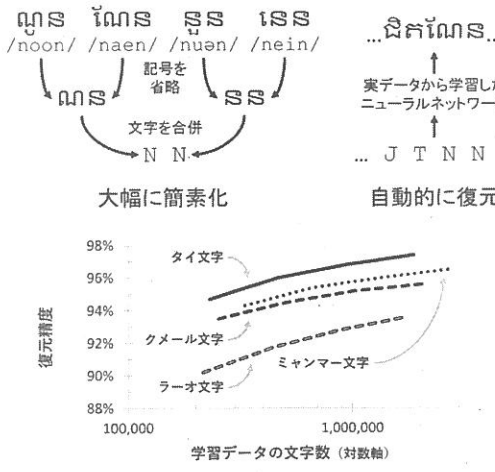
一般的に文字は言語を記録する符号のことである。世界中の言語は数千種類もあり、それらを記録する文字体系も異なる。例えば和文に使われる仮名は「音節文字」を言い、子音・母音ごとに記号を設けるローマ字は「音素文字」を言う。音節・音素文字はともに「表音文字」である一方、漢字のような「表語文字」の体系もある。欧米諸言語に音素文字が採用されることに対し、アジアには複雑な文字体系が用いられる。日本では、進んだ和文入力技術がすでに日常生活に浸透している。

一方、人口が多く経済発展が目覚ましいアジアに広く使われる「アブギダ系文字」の処理は、十分に発展していない。アブギダ系文字は「音素音節文字」とも言う。仮名のような「独立文字」があり、母音の変換・子音の結合を表す「付加記号」もある。従来の入力方式では、文字・記号の順番を正しく組み合わせないと文字化けが生じる。さらに文字と記号の種類数が多いため、キーボード上の配置が洗練されないところも問題点となる。

私が考案した入力方式は、付加記号を省略し、読みが同一または類似である独立文字を区別しないことにする。これにより元のつづりが大幅に簡素化できる。そして、最先端のニューラルネットワーク技術を駆使し、大量の実データから「復元モデル」を訓練する。簡素化されたつづりに自動的・高精度の復元を実現した。

図の上部はカンボジアのクメール文字を例として、簡素化・復元のイメージを示している。下部はタイ文字・ミャンマー文字・クメール文字・ラーオ文字における復元精度を示している。利用される実データ量の増加に従い、復元性能が向上する。現在、私たちは、実用化に向け、入力ソフトウエアを開発し、アジア言語処理技術の向上を目指している。

（火曜日に掲載）



アブギダ系文字の簡素化および自動復元