応般

［シニアコラム］

IT好き放題

［No.36］

捨てる技術

長尾　真 Makoto NAGAO

［名誉会員］
1959年京大電子卒業，1973年京大教授，1997年より6年間京大総長，その後，情報通信研究機構初代理事長，国立国会図書館長を歴任．文化功労者．元本会会長．

私は長年機械翻訳の研究をしてきた．研究の初期には文法による文の解析と翻訳という方法でいろんな成果を得たが，文の微妙な表現の翻訳については必ずしも良い成果が得られない限界というものを強く感じるようなところまで行き，壁にぶつかった．つまり言語の骨格となる理論的枠組みでやれることには限界があり，現実世界の個々の持つ表現の微妙さまでを扱って翻訳をする，いわば第3次近似の世界を扱うには，第2次近似である言語の理論的枠組みではダメだということであった．

そこで1980年代に入って用例翻訳という方法を提唱した．これはある1つの表現が他言語でどのように表現されているかという用例をたくさん集め，与えられた文をそれらの用例の翻訳を頼りに訳すという方式である．しかしこの方式は理論的でないとして1990年代の中ごろまでほとんど無視されていた．ただ世界各地の“理論”に基づいた翻訳研究が行き詰まった結果，用例翻訳の良さが認識され徐々に広まってきている．この考え方は今日では画像の認識にも取り入れられてきているが，ほかにも使える技術である．

用例翻訳を高い質で実現しようとすると，数百万以上の表現について対訳データベースを用意しなければならないが，これを作るためには数億，数十億というテキストのビッグデータを解析する必要がある．そしてこれが今日データ量や処理速度の上でも可能となってきたのである．私が研究していた1960～80年代には考えられないことであった．

ビッグデータといえば，グーグルは世界中のネット上の情報を刻々と集め，整理し，検索の対象として提供しているが，その情報量は今日膨大となり，数えきれない数の記憶装置を並べた巨大記憶工場を設けている．そこに供給する電力は発電所を隣に設けないと賄えないという．情報の量は無限に増えてゆくからグーグルのネット情報の収集はあと10年，20年したらどうなるのか，考えるだけでも恐ろしい．

そこで我々は情報を捨てる技術を開発しなければならなくなる．考えられる1つの確実な方法は次のようなものだろう．すなわち，理論的枠組みで説明できる実例はその枠組みと実例を作り出すパラメータだけを記憶し，実例はすべて捨てる．こうしても“理論＋パラメータ”で実例が復元できるから構わない．そしてこの枠組みに入らないものを例外として残すというわけである．比較的素直なデータ集合では例外的なものは10%にもならないだろうから，これでデータ量はほぼ10分の1になる．また類似のものを1つに吸収するという考え方を導入すれば，例外は1%以下になるだろう．そうすれば100分の1程度になる．こうすればあと1年でパンクする記憶工場が10年，100年と延びることになる．

もう1つは，20年，30年経っても一度も使われなかったデータは捨てるといったことだろう．研究論文などにあてはまるのだろうか？　先祖のお墓も徐々に整理しないと新しい人の入るところがなくなってしまうわけである．長く残るのは立派な人のお墓か，そうするとビッグデータも例外こそ大切であることが類推される．

いずれにしてもこれからはビッグデータに存在する理論的枠組みを明らかにするだけでなく，これに基づいて情報を捨てる研究をする時代に来ているのではないだろうか．

（2013年9月4日受付）