DX 今物語 ⑮

NTTアドバンステクノロジ
AIロボティクス事業本部
IAイノベーションビジネスユニット
主幹担当部長
松尾 義博

◇まつお よしひろ NTT研究所にて自然言語処理の研究開発に従事。日本語辞書・解析、機械翻訳、ウェブ解析、音声対話技術などに取り組む。現在は文書DXの実現に向け、シソーラス辞書および文書理解エンジンの開発開発を担当する。

DXを阻むコトバのゆらぎ

コロナ禍で大きく変わったもののひとつに業務のペーパレス化が挙げられます。それまでも森林資源の保護は大きく叫ばれ、PPC用紙使用量などの環境目標を設定していた組織も多かったと思いますが、お客様都合での会議資料など、なかなか減らせない部分もあちこちに残っていました。

それがコロナ禍で業務スタイルが一変。勤務はリモート、会議も遠隔開催、したがって資料も当然電子ファイルといった習慣がすっかり定着してきました。

しかし、「ペーパー（紙）」はなくなりましたが「ペーパー（書類・文書）」はまだまだ健在です。

印刷されていないだけで、見積書や稟議書、会議資料などはこれまでどおりに作成され電子的に流通しています。これらの電子文書は、作成から保管・廃棄まで、そのライフサイクルすべてがデジタルデータとして取り扱われていますので、一見DXとの相性は良さそうに思えますが、文書をしっかりと活用しようとするとコトバの多様性の問題に遭遇します。

人間のコトバ＝自然言語は、同じことを意味する表現が実に多様で、人が見ると自明なものでもコンピューターには判定困難なものが数多くあります。弊社「NTTアドバンステクノロジ」は「NTTーAT」と略すことが多く、また、登記名は「エヌ・ティ・ティ・〜」と「NTT」部分がカタカナ表記になります。身近なコトバでもたとえば「引換券」は送り仮名ゆらぎの「引き換え券」や、別漢字の「引替券」などが使われます。「クーポン」や「バウチャー」という単語で同一物を指すこともあるでしょう。

こういった表記ゆれや同義語は、文書の検索や分析、テキスト処理の自動化の際に問題となります。同義関係をコンピューターが理解できないと、「引換券」のキーワードでは「クーポン」と書かれた文書は検索できません。

キーワードに基づいて自動応答するチャットボットシステムでは、可能性のあるキーワードを応答ルールに列挙するといった手間が生じるでしょう。

顧客名簿の整理といったシーンも考えられます。お問い合わせなどでお客様に所属名を記入いただく場合、登記法人名を記載してもらえるとは限らず、通称や略称が記入されることも多々あります。これら名簿を名寄せしようとすると表記ゆれは課題となります。

これらのコトバのゆらぎを多数収録したものに「シソーラス」があります。シソーラスとは意味の同一性、類似性などに基づいてコトバを整理した辞書で、システムに導入することで同義性の自動判定が可能になります。検索ヒット率やテキスト情報整理にお悩みの際には、シソーラス辞書の導入を一考ください。