سومین کنفرانس فناوری اطلاعات و دانش
۶ تا ۸ آذرماه ۱۳۸۶

بازشناسی متن چاپی فارسی بر مبنای جداسازی هوشمند

حسین خسروی [*] [**]، احسان الله کبیر [*]

* بخش مهندسی برق، دانشگاه تربیت مدرس
** واحد تحقیق و توسعهٔ شرکت هدی سیستم

E-mail: hosseinkhosravi@modares.ac.ir, kabir@modares.ac.ir

چکیده - یک روش سریع و دقیق برای بازشناسی متن چاپی فارسی با درجهٔ تفکیک 300 نقطه بر اینچ معرفی می شود. این روش مبتنی بر جداسازی زیرکلمات به حروف و زیر حروف سازندهٔ آنها بوده و فرایند بازشناسی در چندین مرحله، با استفاده از طبقه بندهای شبکهٔ عصبی تقویت شده انجام می گیرد. جداسازی زیرکلمات، همواره یکی از مشکل ترین بخشهای بازشناسی متون فارسی و عربی بوده است. کمترین اشتباه در فرایند جداسازی، موجب گسترش خطا در فرایند کلی بازشناسی می شود. در این مقاله علاوه بر ارائهٔ روش ساده و سریع برای جداسازی، با استفاده از نتایج مرحلهٔ بازشناسی، خطاهای مرحلهٔ جداسازی تصحیح می شود. به عبارتی، سیستم دارای یک حلقهٔ بازخورد است که باعث افزایش قابلیت اعتماد آن شده است. داده های هدف در این تحقیق، متون فارسی با قلمهای لوتوس، نازنین و میترا بوده است. البته الگوریتم به گونه ایست که برای سایر قلمها قابل توسعه است. این روش روی 8 صفحه متن فارسی با درجهٔ تفکیک 300 نقطه بر اینچ آزمایش شده و دقت بازشناسی 99% حاصل شده است.

کلید واژه –جداسازی، بازشناسی متن، فارسی

1- مقدمه

بازشناسی متن، یکی از زمینه هایی است که در چند دههٔ اخیر، فعالیتهای زیادی را به خود اختصاص داده است [1-7]. پژوهشگران متعددی روی روشهای مختلف بازشناسی متن کار کرده‌اند و امروزه سیستمهای تجاری با دقت بالا برای بسیاری از زبانهای دنیا وجود دارد. بازشناسی متن یکی از مهمترین بخشهای دولت الکترونیک نیز به شمار می رود و در سالهای اخیر، در کشور ما تقاضا برای یک سیستم بازشناسی متن فارسی، خاصه در سازمانهای دولتی، به شدت افزایش یافته است. هر چند در ایران تحقیقات متعددی در زمینهٔ بازشناسی متن گسسته [8-12] و پیوسته [4, 13-15] صورت گرفته است، لیکن به دلیل عدم هماهنگی و عدم پشتیبانی مناسب، هنوز یک سیستم تجاری قابل اعتماد برای بازشناسی متن پیوستهٔ فارسی نداریم.

روشهای بازشناسی متن فارسی، به دو دستهٔ کلی تقسیم می‌شوند [16]: بازشناسی بر مبنای شکل کلی و بازشناسی بر پایهٔ جداسازی. در روش اول، کل زیرکلمه به عنوان یک شکل در نظر گرفته شده و تشخیص داده می شود و در روش دوم، ابتدا زیر کلمهٔ ورودی به حروف و زیر حروف تجزیه شده و سپس بازشناسی صورت می گیرد. گاهی اوقات نیز از ترکیب این دو روش استفاده شده است. در این مقاله ما با استفاده از روشی مبتنی بر جداسازی با حلقهٔ بازخورد و همچنین استفاده از دانشهای جانبی مانند نقاط، علائم و برخی قواعد گرامری ساده به بازشناسی متن فارسی پرداخته ایم.

در زمینهٔ جداسازی نوشته های فارسی و عربی، روشهای متعددی معرفی شده است که می توان آنها را در چهار دسته تقسیم بندی کرد: روشهای مبتنی بر افکنش، روشهای مبتنی بر کانتور، روشهای مبتنی بر پروفایل و روشهای مبتنی بر اسکلت.