OCR: آموزش الفباي فارسي به رايانه

رايانه تنها يك ابـزار اسـت، و مـا انسـان‌هـا مـي‌بايـد شـيوة كـار كـردن را بـه او بيـاموزيم، و امكانات كار را هم بـرايش فـراهم آوريـم، نـرم‌افزارهـاي مختلـف بـه همـين منظـور طراحـي مي‌شوند. يكي از آنها نرم‌افزار OCR است كه براي تشخيص و بازيابي الفبـا (نوشـته‌هـاي دسـت‌نويس يا تايپ‌شده) توسط كامپيوتر طراحي مي‌شوند.
اگر هر يك از ما به طور متوسط ۷۰ تا ۸۰ سال عمـر كنـيم مجمـوع روزهـاي زنـدگيمان بـين ۲۵هزار تا ۲۹هزار روز مي‌شود. حال اگر به ما بگويند نرم‌افزاري توليد مي‌شود كه مي‌تواند بسيار بيشتر از عمر چندين و چند انـسان در وقـت صـرفه‌جـويي كنـد حتمـاً از آن اسـتقبال مي‌كنيم، مگر آنكه براي چند صدبرابر روزهاي عمرمان ارزشي قائل نباشيم!
نرم‌افزار OCR قادر است چنين تحولي را در استفاده وقت بشر به وجود آورد. ايـن نـرم‌افـزار مي‌تواند متن‌هاي دست‌نويس يا متوني را كه قبلاً تايپ شده‌اند و اكنـون فايـل تـايپي آنهـا موجود نيست، خود به تنهايي و بدون دخالت انگـشتان هـيچ تايپيـستي تايـپ كنـد. فـرض كنيد كه مثلاً مي‌خواهيم متن مقالات روزنامه اطلاعات سال ۱۳٤۰ شمسي را (كه اكنـون نه تنها فايل تايپي‌اش موجود نيست ـ چون آن زمان اصلاً تايپ كـامپيوتري در كـار نبـود! ــ بلكه خود نسخه‌هاي روزنامه را هم به زحمت مي‌توان پيدا كـرد) تايـپ ديجيتـالي كنـيم، و اين متن‌ها را داخل بسته‌هاي نرم‌افزاري يا اينترنت قرار دهيم. اگر هر شـماره از روزنامـه را ۲٤ صفحه فرض كنيم، و هر تايپيـست بتوانـد در هـر روز حـداكثر يـك صـفحه از آن صـفحات كاهي و كهنه شدة قديمي را دوباره تايپ كند، مجموعاً ۲٤ روز لازم است تـا تنهـا مقـالات يك شماره از روزنامه تايپ شود. بنابراين در عرض يك سال يك نفر مي‌تواند تنها ۱۵ شماره از روزنامه را تايپ كند. حال اگر نـرم‌افـزاري باشـيم كـه بتوانـد بـا اسـكن كـردن هـر صـفحة روزنامه، به طور خودكار مقالات آن را تايپ كند، تحولي عظيم رخ مي‌دهد، يعنـي مطالـب و مقالات هزاران شماره از روزنامه‌هاي قديمي به سرعت وارد فايل‌هاي رايانه‌اي مي‌شود.
حال اين امكان را تعميم بدهيد به هزاران كتاب و دستنويس‌هاي قديمي يـا جديـد، كـه هـر كس بخواهد تنها يك صفحه از آنها را تايپ كند، بايد كلـي وقـت صـرف كنـد. مـي‌بينيـد كـه نرم‌افـزار OCR براسـتي مـي‌توانـد هـزاران هـزار روز در وقـت مـا صـرفه‌جـويي كنـد، و البتـه هزينه‌ها را هم كاهش دهد. البته فقط يك مشكل كوچك به وجود مي‌آيد و آن بيكـار شـدن تايپيست‌هاست!

OCR در ايران چگونه آغاز شد؟

امان از دست اين تيزهوشـان! مـاجرا از ثبـت‌نـام داوطلبـان آزمـون «سـازمان ملـي پـرورش استعدادهاي درخشان (تيزهوشان)» در سال ۱۳۸۰ آغاز شد. ثبت‌نام از روي فرم‌هايي كه توسط دانش‌آموزان تكميل مي‌شد انجام مي‌گرفت. دانش‌آموزان شركت‌كننده در آزمـون ــ مانند آزمون‌هاي سراسري ـ بايد نام، نام خانوادگي، نام پدر، نام شهرسـتان محـل تولـد و سكونت، نام مدرسه و دين خـود را در داخـل كادرهـاي مربعـي شـكل و بـه صـورت حـروف مقطع (يعني هر حرف داخل يك كادر) مي‌نوشتند. وقتي كه همة فرم‌ها از طريق پست به سازمان مركزي برگزاركننده آزمون مي‌رسيد، عدة زيادي تايپيست مـتن آنهـا را دوبـاره وارد رايانه مي‌كردند. در واقع همان حرف‌هاي داخل كادر را دوباره تايـپ مـي‌كردنـد تـا اطلاعـات شناسنامه‌اي هر دانش‌آموز به صورت ديجيتالي درآيد. اين روش هـم بـسيار زمـان‌بُـر بـود و هم نياز به تعداد زيادي تايپيست داشت. احتمال داشت كه تايپيسـت‌هـا هـم هنگـام تايـپ اشتباه كنند و با ثبت نادرست يك نام، مشخصات فردي در رايانه مركزي وارد شود كه اصلاً متولد نشده است! مثلاً فرض كنيد تايپيست محترم نـام «جـواد» را، كـه داخـل كادرهـا بـه صورت «ج.و.ا.د» نوشته شده بود، «فؤاد» تايپ مي‌كرد؛ در آن صـورت در كـارت شناسـايي

جواد سابق، فؤاد فعلي ثبت مي‌شـد! (جـوادِ موجـود حـذف مـي‌شـد و فـؤاد نـاموجود وارد فهرست داوطلبان مي‌شد!) افزون بر اين، هزينة كار نيز بسيار زياد بود.
به علت همين مشكلات، در بهمن‌ماه ۱۳۸۰، نخستين طرح OCR براي بازشناسي حـروف فارسـي توسـط كـامپيوتر ارائـه شـد و در سـال‌هـاي ۱۳۸۱ و ۱۳۸۲ نيـز ثبـت‌نـام آزمـون تيزهوشان به ياري اين نرم‌افزار انجام شد.

OCR چيست؟
OCR سرنام اصطلاحي است كه صورت كامل آن در واژه‌نامه‌هـاي انگليـسي بـه دو صـورت آمده است:
۱. Recognition Optical Character
۲. Optical Character Reader
فرض كنيد كـه مـا متنـي را روي كاغـذ داريـم و مـي‌خـواهيم آن را وارد رايانـه كنـيم. اولـين روشي كه به ذهن مي‌رسد اين است كه متن را به تايپيست بـدهيم تـا بـا كـامپيوتر تايـپ كند. اما آيا مي‌شود عين همان متن را وارد رايانـه بكنـيم تـا نيـازي بـه تايـپ نباشـد؟ البتـه دستگاه «اسكنر» مي‌تواند تصويري از آن متن را وارد رايانه كند، تا اينجا بخشي از مـشكل ما حل شده است. اما رايانه كه نه عقلي دارد و نه «زبان» مي‌فهمد، نمـي‌توانـد حـروف و كلمـات را از هـم تـشخيص دهـد. مـثلاً اگـر از كـامپيوتر بخـواهيم بـه مـا بگويـد كـه در مـتن اسكن‌شده كلمة «علي» چند بار آمده است، بي‌آنكه شرمنده شود، مي‌گويـد: «error»، يعني: «نمي‌توانم تشخيص بدهم!» در واقع ايـن «تـصوير ديجيتـال‌شـده» بايـد بـه «تـصوير قابل پردازش» تبديل شود. موضوع اصلي OCR همين است.

انواع OCR
در زبان‌هاي ديگر، به ويژه زبان‌هايي كه با حروف لاتيني نوشته مي‌شوند، سال‌هاست كه از OCR استفاده مي‌شود. اما در ايران تازه دو سـه سـالي اسـت كـه بـه فكـر اسـتفاده از OCR در زبان فارسـي افتاده‌ايم.
و اما OCR چند نوع است: يا تايپي است يا دست‌نويس. يعني يا بايـد يـك مـتن قـبلاً تايـپ شـده را (مثل كتاب‌ها و روزنامه‌هاي چندين سال قبل، يا حتي متني را كـه فايـل تـايپي آن موجود نيست و فقط پرينت آن را داريم) وارد رايانه كنيم، يا متن دست‌نـويس را. مـتن‌هـاي دست‌نويس هم به دو صـورت «گسـسته» و «پيوسـته» وجـود دارنـد: مـتن «دسـت‌نـويس پيوسته» مثل همان چيزهايي است كه ما هرازگاهي كه دلمان تنگ مـي‌شـود روي كاغـذ مي‌نويسيم، يا يك نامه، يا يك قطعه شعر و ... امـا مـتن «دسـت‌نـويس گسـسته» همـان نوشته‌هايي است كه حروف آن جدا از هم و به صورت گسسته نوشته شده‌اند، مثـل نـام و نام‌خانوادگي كه در فرم‌هاي آزمون ثبت‌نام، به صورت هـر حـرف داخـل يـك كـادر، نوشـته مي‌شوند. طراحي OCR گسستة فارسي تقريباً در مراحل پاياني كار قرار دارد ولـي، OCR پيوسته ظاهراً سال‌هاي زيادي كار مي‌برد.
«رضا صـديق» و «پرويـز رزازي»، كـه در رشـتة مخـابرات تحـصيل كـرده‌انـد و مـسئولان يـك شركت كامپيوتري به نام «انديشه نرم‌افزار پايا» هستند، براي اولين بار به طور جدي پروژة OCR فارسـي را دنبال كـرده‌انـد. رزازي كـه دانـشجوي مخـابرات و مـسئول بخـش پـردازش سيگنال شركت «پايا» و مدير پروژة OCR در اين شركت است، مـي‌گويـد : « OCR در دنيـا موضـوعي ناشـناخته نيـست، و بـر روي آن زيـاد كـار شـده اسـت، ولـي در ايـران بـا آنكـه مدت‌هاسـت روي آن كـار شـده، امـا بـسياري از ايـن كارهـا در حـد كارهـاي دانـشگاهي و مقاله‌هاي علمي باقي‌مانده بـود و تبـديل بـه يـك محـصول كـاربردي در ابعـاد وسـيع (مثـل ثبت‌نام آزمون‌هاي بزرگ) نشـده بود. ما بر روي اين طرح كار كرديم و هدفمان هـم ايـن بـود كه محصول را به شكل صنعتي آن توليد كنيم.»

البته غير از شركت «پايا»، دو شركت ديگر نيز با حمايت دبيرخانه طرح «تكفا» (توسعه كاربرد فناوري اطلاعات و ارتباطات) مشغول پژوهش و آزمايش بر روي OCR فارسي هستند. يكي از اين شركت‌ها «داده‌پردازان دوران نوين» نام دارد كه مديريت آن را دكتر «حسام فيلي» بر عهده دارد. دكتر فيلي متخصص در رشتة هوش مصنوعي، از دانشگاه صنعتي شريف، است و شركت «دوران نوين» را از سال ۱۳۸۱، با هدف كار تخصصي بر روي پروژه‌هاي هوش مصنوعي تأسيس كرده است. او دربارة چگونگي پيوستن شركتش به اين طرح مي‌گويد: «از تيرماه سال ۸۲ با شروع فعاليت طرح «تكفا» و حمايت‌هاي مالي آنها، اين شركت تصميم گرفت كه در زمينة طراحي OCR فارسي پژوهش و فعاليت كند. اين پروژه در شركت «دوران نوين» با همكاري آقاي «دكتر ابراهيمي مقدم» كه او هم از دانشجويان دورة دكتري هوش مصنوعي دانشگاه صنعتي شريف است، انجام مي‌گيرد.

فارسي ما و مشكلات آن

قبل از اينكه به مراحل ديگر OCR بپردازيم، لازم است اندكي هم به مشكلات خط فارسي ـ يا در واقع ويژگي‌هاي اين خط ـ بپردازيم. اول اينكه ما در فارسي حروف را به صورت چسبيده و پيوسته مي‌نويسيم و اين كار براي تشخيص حرف به حرف نوشته از سوي رايانه (كه قرار است در مراحل بعدي آن را تايپ كند) بسيار مشكل است. تصور كنيد كه همين كلمه ساده «است» را به حالت‌هاي مختلف مي‌شود نوشت: يكي براي «س» دندانه مي‌گذارد، يكي نمي‌گذارد، يكي آن را مي‌كشد و يكي نمي‌كشد و... حالا اگر همين صورت‌هاي مختلف «س» به «ت» هم بچسبند، تشخيص حروف براي ما انسان‌ها هم سخت مي‌شود، چه رسد به رايانه.

شباهت حروف

مشكل ديگر خط ما اين است كه حرف‌هاي فارسي بسيار به هم شبيه‌اند. مثلاً در نظر بگيريد كه تفاوت «ر» با «ز» با «ذ» يا «ب» با «ت» تنها در يك نقطه است، و چون نقطه جزء بسيار كوچكي است، اگر يك خط يا حتي يك لك كوچك روي كاغذ بيفتد، تشخيص حروف از هم بسيار دشوار مي‌شود و دردسر جدي براي بازشناسي حروف توسط رايانه ايجاد مي‌كند. اينها تازه مشكلات خط فارسي است. دربارة اعداد فارسي هم اين مشكل وجود دارد: صفر ما يك نقطه كوچك است كه مي‌تواند رايانه را به اشتباه بيندازد؛ اعداد ٤، ٣، ٢، ١ هم بسيار به هم شبيه هستند و تنها تفاوتشان يك دندانه كوچك است.

به دلايل گفته شده OCR درمرحلة كنوني در كشور ما مربوط به «دست‌نويس‌هاي گسسته» يا متن‌هاي تايپي پيوسته است، و تا بازشناسي متن‌هاي دست‌نويس پيوسته توسط كامپيوتر راه زيادي در پيش است، چون در دست‌نويس‌هاي گسسته، اگرچه حروف به هم شباهت دارند، حداقل جداجدا نوشته شده‌اند. در متن‌هاي پيوسته تايپي هم مشكل كشيده شدن يك حرف يا شكسته نوشته شدن حروف را نداريم. البته به گفته مسئولان شركت «پايا» در حال حاضر هم نرم‌افزارهايي وجود دارد كه متن دست‌نويس پيوسته را تبديل به حروف جدا ازهم و گسسته مي‌كنند، ولي ضريب خطاي اين نرم‌افزارها زياد است و به شكل صنعتي درنيامده‌اند.

بازشناسي حروف و الگو

تا اينجا گفتيم تصوير صفحه‌اي كه در آن حروف به طور جداجدا (هر حرف داخل يك كادر) نوشته شده است، به وسيلة اسكن وارد رايانه مي‌شود. مرحلة بعدي اين است كه حروف بازشناسي شوند، يعني مكان آنها از ديگر خطوط (مثل خطوط كادري كه داخل آن نوشته شده) بازشناسي شود، و اگر متن پيوسته تايپي است، حروف جدا شوند و زوايد تصوير حذف شود. مثلاً اگر دانش‌آموزي «س» را به گونه‌اي نوشت كه بيرون از كادر بود، به رايانه بفهمانيم كه بي‌دقتي شده است او بايد همان حرف داخل كادر را بخواند. در

مرحلة بعدي كه «بازشناسي الگو» نام دارد، با تعدادي شرط مي‌شود فهميد كه مثلاً حرفي «الف» است يا نه، و رايانه تشخيص مي‌دهد كه حرف «پ» است يا «ب». براي اين تشخيص لازم است كه تصوير حرف «الف» با الف‌هاي نمونه ـ كه قبلاً به رايانه داده شده است ـ منطبق شود. الفباي نمونه قبلاً از روي يك مجموعه بزرگ آموزشي تهيه شده و ويژگي‌هاي مشترك از آن استخراج شده است. اما از آنجا كه تنوع صورت‌ها نوشتاري يك حرف به صورت دست‌نويس بسيار زياد است، مدلي آماري استخراج مي‌شود كه در آن شباهت ويژگي‌هاي استخراج شدة قبلي با نمونه ورودي به رايانه بررسي مي‌شود. در اينجا «بازشناسي الگو» با روش‌هاي آماري انجام مي‌شود كه روش معمول در سيستم‌هاي OCR است.
اگر فكر مي‌كنيد كه كار تمام شده است در اشتباهيد، چون تازه مي‌رسيم به دنبالة حروف. مثلاً اگر كسي همان حرف «س» را با دنباله بنويسد، رايانه بايد تشخيص دهد كه اين حرف فقط «س» است، يا مثلاً «ي» هم به آن چسبيده است.

مدل‌سازي يا پردازش زباني
مرحله بعدي «مدل‌سازي زباني» يا «پردازش زباني» نام دارد. حروف به هم چسبيده، كه كلمه را درست مي‌كنند، بايد معني‌دار يا شناخته‌شده باشند. در اين مرحله بررسي مي‌شود كه چه كلماتي در زبان وجود دارد؟ چه تركيب‌هايي از كلمات مجاز است؟ و... البته در مراحل پيشرفته‌تر، مدل‌سازي گرامري (دستور زبان) و مدل‌سازي معنايي هم وجود دارد كه تشخيص مي‌دهد جمله از لحاظ دستوري و معنايي درست است يا بي‌مفهوم است. اما در OCR گسسته ـ كه بيشتر براي ثبت‌نام استفاده شده ـ شباهت يك كلمه به نام، نام خانوادگي، شهر و ... كافي است.
براي تشخيص تركيب‌هاي مجاز يك كلمه يا معني‌دار بودن يك كلمه نيز به تهية بانك‌هاي اطلاعاتي (Data base) نياز داريم. در اين بانك‌ها مثلاً تمام نام‌هاي كوچك و بزرگ ايرانيان قبلاً جمع‌آوري شده است و هنگام تطبيق يك كلمه با آن مشخص مي‌شود كه رايانه حروف آن را دست تشخيص داده يا نه. بنابراين نقش اين بانك اطلاعاتي بسيار مهم است، چون اگر نامي در آن ثبت نشده باشد، كلمه‌اي كه آن نام را شامل شود، به طور خودكار از برنامه OCR حذف مي‌شود يا پيغام مي‌آيد كه: «اين كلمه اشتباه است» در صورتي كه ممكن است مثلاً نام «هشام» در بين نام‌هاي ايراني وجود داشته باشد، ولي قبلاً در بانك اطلاعاتي ثبت نشده باشد.

بانك‌هاي ما و ديگران
مهندس «رزازي» دربارة مشكل بانك‌هاي اطلاعاتي در زبان فارسي مي‌گويد: «در دنيا براي توسعة OCR و ارزيابي آن، بانك‌هاي اطلاعاتي استاندارد ساخته شده است كه در آنها همة كلمات وجود دارند، يعني بانك هم مشكل ديجيتال كلمه را دارد، و هم تصويرش را. اما براي زبان فارسي، اين بانك‌هاي اطلاعاتي چه براي ارزيابي و چه براي توسعه، استاندارد شده نيست. در واقع هر كسي براي خودش يك بانك اطلاعاتي مي‌سازد، و اين نمونه‌هاي متفاوت مشكلاتي را ايجاد مي‌كند. مثلاً براي ثبت‌نام دانش‌آموزاني كه در آزمون مدارس تيزهوشان شركت كرده بودند، يك بانك اطلاعاتي حاوي نام‌هاي فارسي، از روي اطلاعات فرم‌هاي سال‌هاي قبل، تهيه شد كه از روي آن كلماتي كه خيلي شبيه به نام‌هاي فارسي بودند تشخيص داده مي‌شد. مثلاً اگر رايانه كلمه‌اي را «مصيبت» تشخيص داد، براساس بانك اطلاعاتي معلوم مي‌شود كه «مصيب» بوده است كه يك نام ايراني است.

علي، ولي، قلي ... و سيب
نكتة ديگر اين است كه يك بانك اطلاعاتي بايد شامل تعدادي كلمات خام باشد، بلكه «بسامد» آن واژگان، يعني ميزان استعمال و تكرار كلمات در زبان و مشخصات آماري آن‌ها هم بايد ثبت شده باشد، والا كارايي زيادي ندارد. مثلاً «علي» نامي است كه شباهت زيادي به «ولي» و «قلي» دارد. كارهاي آماري در بانك اطلاعاتي بايد طوري انجام شده باشد كه تعداد «علي» بيشتر باشد، تا و بعد نوبت «ولي» و «قلي» برسد، چون درصد بسامدي «علي» به لحاظ آماري و كاربرد در ميان نام‌هاي بيشتر است.
در OCR فارسي گسسته، اگر فقط مربوط به نام‌ها و نام‌خانوادگي باشد، كار ساده‌تر است از حالتي كه در OCR پيوسته وجود دارد. چون در OCR پيوسته هر كلمه‌اي ممكن است وجود داشته باشد مثل «سيب»، اما در بانك اطلاعاتي نام‌ها همه مي‌دانيم كه سيب نام يك شخص نيست بلكه نام يك ميوه است! بنابراين در OCR همواره سعي مي‌شود كه درصد خطا كاهش يابد، تا كلمات در حد ممكن درست تشخيص داده شوند. اگرچه طراحان هنوز به صددرصد صحت نرسيده‌اند، ولي نگران نتايج آزمون خود نباشيد، چون تمامي اطلاعات مربوط به شما چندين بار كنترل مي‌شوند و از سازوكار reject (يا مردودي) در رايانه هم استفاده مي‌شود. در اين روش اگر رايانه نتوانست كلمه‌اي را تشخيص دهد، مي‌فهمد كه نفهميده است و در خروجي‌اش مي‌آورد كه: «من اين كلمه را نفهميده‌ام» و كار به سيستم دستي مي‌رود و در آنجا تصحيح مي‌شود. اين فرايند در پست خيلي كارايي دارد. در هر جاي دنيا كه تفكيك نامه‌ها و ديگر مرسولات پستي به وسيلة OCR انجام مي‌شود، بعضي از نامه‌ها در سيستم كامپيوتري وارد سازوكار «مردودي» مي‌شوند و به طور دستي مورد بررسي مجدد قرار مي‌گيرند. هم‌اكنون در سطح محدودي از OCR در پست كشور ما نيز استفاده مي‌شود، چون در پست هم كد پستي چندرقمي و ديگر اطلاعات به صورت گسسته و داخل كادرهايي نوشته مي‌شود، و كار آسان‌تر است.

در مورد خطاي OCR در تشخيص كلمات، مسئولان شركت «پايا» نظر جالبي دارند:

«حتي با تعبية سيستم مردودي (reject) هم ممكن است خطايي در تشخيص كلمات وجود داشته باشد. بايد در نظر داشته باشيم كه هيچ سيستم پردازشگري (از جمله انسان) بدون خطا نيست. نكته مهم اين است كه يك سيستم ماشيني درصد خطاي كمتري نسبت به انسان داشته باشد تا جايگزين خوبي براي انسان باشد. مسئله اين نيست كه خطا را به صفر برسانيم. هر قدر كه فناوري جلوتر مي‌رود، ميزان خطا هم بيشتر كاهش مي‌يابد.»
مدير شركت «دوران نوين» هم به گونه‌اي ديگر به همين موضوع اشاره مي‌كند: «انتظار ما از مسئولان طرح «تكفا» آن است كه با موضوع OCR واقع‌بينانه‌تر برخورد شود، و در بحث مربوط به هزينه‌هاي پروژه و انتظاراتي كه از OCR مي‌رود، واقعيت‌ها در نظر گرفته شود. ديدگاه كنوني مسئولان تكفا آن است كه كل مشكل «خطا» تا ۱۰۰درصد حل شود، در حالي كه فكر مي‌كنم حل مسائل مربوط به هوش مصنوعي نياز به روش تدريجي دارد. مثلاً در زبان عربي هم، نرم‌افزار «صخر» در نسخة اول خود فقط تا حدود ٤٠درصد دقت داشت، در حالي كه اكنون پس از گذشت ۱۳ سال از اولين نسخه آن دقت به مرز ۹۸ درصد رسيده است.»
همان‌طور كه اشاره شد از OCR در ثبت‌نام آزمون «سازمان ملي استعدادهاي درخشان» در سال‌هاي ۸۱ و ۸۲ استفاده شد كه از طريق آن ٤٤٠،٠٠٠ نفر به طور ماشيني ثبت‌نام شدند. اين روش باعث شد كه در سال ۸۱ (نمونه اول) ٤٥ درصد در هزينه‌ها و ۲۵ درصد در زمان ثبت‌نام صرفه‌جويي شود. در سال بعد (۸۲) اين رقم به ۵۰ درصد رسيد.

نرم‌افزاري كه در اين آزمون‌ها مورد استفاده قرار گرفت براي هر كدام از موارد صحت بازشناسي متفاوتي داشت و در مجموع كار آن خوب بود.
به نظر مي‌آيد كه در چند سال آينده و با پيشرفت OCR فارسي و كاهش هر چه بيشتر خطاي آن، در آزمون‌هاي بزرگ‌تري مانند آزمون سراسري دانشگاه‌ها نيز بتوان از آن استفاده كرد.

سرنوشت OCR دست‌نويس
در مورد OCR پيوسته دست‌نويس نيز روند كار به همان صورتي است كه شرح داديم، اما آنچه كار را دشوارتر مي‌كند، قطعه‌بندي و جداجدا كردن حروف به هم چسبيده و تشخيص آنهاست. اگر اين روند طي شود، اين اميد وجود دارد كه روزي از OCR پيوسته دست‌نويس فارسي هم در سطح گسترده‌اي استفاده شود. البته OCR پيوسته دست‌نويس حتي در زبان انگليسي هم هنوز به كاربرد وسيع و عملي نرسيده است. مهندس «رزازي» دراين مورد مي‌گويد: « OCRانگليسي در سيستم عامل windows وجود دارد كه همراه با office فروخته مي‌شود، ولي فكر نكنيد كه نامه‌هاي اداري انگليسي كه با دست‌نويس نوشته شده‌اند همه با OCR تايپ مي‌شوند. اين كار براي به نتيجه رسيدن به حداقل يك روند ۱۰ ساله را بايد طي كند. OCR فارسي يك مرحله عقب‌تر است، پس زمان بيشتري مي‌برد.»
مهندس «صديق»، مديرعامل شركت «پايا» هم مي‌گويد: «همين OCR فارسي گسسته هم تا چند سال پيش يك رؤيا بود، ولي ديديم كه محقق شده است و به مرور پيشرفته‌تر هم خواهد شد. بنابراين طراحي OCR پيوسته فارسي هم، اگرچه سال‌ها طول مي‌كشد، ولي مطمئناً به نتيجه خواهد رسيد. اين طرح يك طرح تحقيقاتي است كه در دانشگاه‌ها دنبال مي‌شود و هنوز به يك محصول صنعتي قابل استفاده در سطح كلان و كاربردي براي عموم نرسيده است. ولي در حال حاضر نمونه‌هاي دانشگاهي و آزمايشگاهي آن در داخل كشور وجود دارد و موضوع رساله دكتري برخي از دانشجويان است.»
بنابراين بين ۱۰ تا ۲۰ سال آينده، آن‌گونه كه مسئولان شركت «پايا» مي‌گويند، OCR پيوسته دست‌نويس فارسي هم وارد بازار خواهد شد.
دكتر فيلي هم در پاسخ به اين سؤال كه «آيا طراحي OCR پيوستة فارسي روزي تحقق خواهد يافت؟» پاسخ مي‌دهد: «بله ولي به تدريج.»
به هر حال براساس قرارداد «تكفا» با شركت‌هاي ايراني، تا كمتر از يك ماه ديگر، نسخه‌نهايي (البته نه صددرصد تكميل‌شده) OCR فارسي دست‌نويس گسسته و تايپي پيوسته ارائه خواهد شد. مدير شركت «دوران نوين» در اين مورد مي‌گويد: «پروژة OCR گسسته در مراحل پاياني خود قرار دارد ولي داراي مشكلاتي در تشخيص انواع اسكنرها و انواع فونت‌هاست كه در حال رفع آن هستيم. اين نرم‌افزار در حال حاضر امكان تشخيص فونت‌هاي تايپي فارسي با دقت زياد را دارد، ولي مشكل جدي آن است كه با اسكنرهاي مختلف نتايج نامناسبي مي‌دهد.» وي از اهميت اين طرح در بعد كلان ملي هم مي‌گويد: «با توجه به اين كه مشكل OCR براي بسياري از زبان‌هاي دنيا مانند انگليسي عملاً حل شده است، اگر در كشور ما هم به نتيجة نهايي برسد در افزايش سطح اطلاعات فارسي در دنياي ديجيتالي امروز (از جمله در اينترنت) بسيار اهميت خواهد داشت.

mohammad6347@yahoo.com