بسم الله الرحمن الرحیم

اصول و مبانی خطایابی املایی

یاسر سوری ۸۶۵۲۱۲۵۸

دانشگاه علم و صنعت ایران

souri@comp.iust.ac.ir

چکیده

مسئله‌ی اختراع روش‌هایی برای تشخیص خودکار و تصحیح خطاهای املایی از دهه‌ی ۶۰ میلادی مورد توجه بوده است. از آن زمان تا کنون تلاش‌های زیادی برای تولید سیستم‌های تشخیص و تصحیح خطاهای خودکار شده است. در این پژوهش به بررسی روش‌های مختلف خطایابی املایی می‌پردازیم.

واژگان کلیدی

پردازش زبان طبیعی، خطایابی املایی، تصحیح خطای املایی

۱. مقدمه

مسئله‌ی اختراع الگوریتم‌ها و روش‌های تصحیح خودکار کلمات در متن به یک چالش بالقوه‌ی تحقیقاتی تبدیل شده است. کار در این زمینه از دهه‌ی ۶۰ میلادی آغاز شده است و تا به حال ادامه داشته است. اگر چه چند نمونه خطایاب املایی خوب تجاری و علمی مدتی است که بوجود آمده است، ولی روش‌های تصحیح کنونی محدود به محدوده و دقت خود هستند [۱].

تفاوتی باید بین کار **تشخیص** خطای املایی و **تصحیح** آن قائل شد. روش‌های کارایی برای تشخیص رشته‌هایی که در یک لغت‌نامه وجود ندارند ایجاد شده است. ولی تصحیح رشته‌ی شامل غلط املایی مسئله‌ی به مراتب سخت‌تری است [۱].

۱.۱ انواع خطایابی املایی

بیش‌تر روش‌های خطایابی املایی موجود بر روی کلمات **منزوی** تمرکز دارند و اطلاعاتی را که از **متنی** که رشته در آن ظاهر شده است می‌تواند به دست آورد را در نظر نمی‌گیرند. چنین روش‌هایی قادر به تشخیص قسمت بزرگی از خطاها از جمله خطاهای چاپی، آوایی و نحوی که منجر به تولید دیگر کلمات مورد قبول هستند، نیستند [۱].

به طور کلی خطایابی املایی را می‌تواند به سه نوع (۱) تشخیص خطای غیر لغت[۱]؛ (۲) تصحیح خطای کلمات منزوی و (۳) تصحیح خطای وابسته به متن؛ دسته بندی کرد.

در این پژوهش به معرفی مختصری از روش‌های خطایابی املایی در این زمینه‌ها خواهیم داشت.

۲. تشخیص خطای غیر لغت

دو روش اصلی که برای تشخیص خطای املایی غیر لغت به کار رفته‌اند، تحلیل چندنگاشت[۲] و مراجعه به لغت‌نامه است. چندنگاشت‌ها زیر دنباله‌های چند حرفی از کلمات و رشته‌ها هستند که چند معمولاً یک، دو و یا سه است. به طور کلی روش‌های تشخیص خطای مبتنی بر چندنگاشتی با امتحان کردن هر چندنگاشت در متن ورودی و به دنبال آن در یک جدول از پیش تولید شده گشتن برای مشخص شدن وجود یا تعداد تکرار آن، کار می‌کند. رشته‌هایی که دارای چندنگاشت‌های بدون تکرار یا تکرار بسیار کم باشند (مثل سه‌نگاشت پخز یا فقث) به عنوان خطاهای احتمالی در نظر گرفته می‌شوند. این روش‌ها معمولاً به یک لغت‌نامه یا یک مجموعه‌ی بزرگ از نوشتجات برای تولید از پیش جدول چندنگاشت‌ها نیاز دارند. روش‌های مبتنی بر لغت‌نامه به سادگی کلمات ورودی را در یک لغت‌نامه جستجو می‌کنند. اگر کلمه‌ی ورودی در لغت‌نامه وجود نداشت به عنوان خطا علامت گذاری می‌شود.

۲.۱. روش‌های تحلیل چندنگاشت

خطاهایی که توسط دست‌گاه‌های تشخیص نوری حروف[۳] معمولاً آن‌هایی هستند که حروف شبیه به هم را با هم اشتباه می‌کنند، مثل الف و لام یا ۱ و الف. روش‌های مبتنی بر تحلیل چندنگاشت برای تشخیص این خطاها خوب عمل می‌کنند.

جدول‌های چندنگاشت می‌توانند به شکل‌های متفاوتی تولید شوند. ساده‌ترین آن‌ها یک دونگاشت دودویی است که یک آرایه‌ی دو بعدی ۳۲×۳۲ است که تمام ترکیب‌های دوتایی حروف را در بر می‌گیرد. هر خانه‌ی آن شامل ۱ اگر آن دونگاشت در متن ما حداقل یکبار تکرار شده باشد یا ۰ است. یک سه‌نگاشت دودویی به همین شکل و یک آرایه‌ی سه بعدی است. هر دو آرایه‌های بالا را چندنگاشت‌های دودویی **غیر وابسته به موقعیت** می‌نامند. بدین معنا که محل چندنگاشت را در کلمه مشخص نمی‌کنند.

در اختصاص فضای بیش‌تر به جدول چند نگاشت می‌توان موقعیت چندنگاشت را هم ذخیره کرد. در آزمایشات مشخص شده است که چندنگاشت‌های وابسته به موقعیت بهتر عمل می‌کنند [۲] [۳].

۲.۲. روش‌های مراجعه به لغت‌نامه

---

[۱] Nonword
[۲] N-gram
[۳] Optical Character Recognition

در مراجعه به لغت‌نامه مسئله زمان پاسخ‌گویی است. معمولاً اندازه‌ی لغت‌نامه‌ها بین ۲۵ هزار تا ۲۵۰ هزار لغت است و برای این اندازه از لغت‌نامه‌ها زمان پاسخ‌گویی بسیار بالا می‌رود. این مشکل از سه روش مراجعه به لغت‌نامه‌ی کارا، تقسیم لغت‌نامه و روش‌های مبتنی بر پردازش ریخت شناسی سعی شده است که حل شود.

پر استفاده‌ترین روش برای دسترسی سریع به لغت‌نامه‌ها استفاده از جدول‌های درهم[۴] است [۴]. برای امتحان کردن یک رشته‌ی ورودی، آدرس درهم آن محاسبه می‌شود و با مراجعه به جدول درهم از پیش ساخته شده و دریافت کلمه‌ی ذخیره شده در آن مکان اگر کلمات یکی نبودند یا کلمه‌ای وجود نداشت، غلط املایی تشخیص داده می‌شود.

حسن اصلی استفاده از جدول‌های درهم عدم نیاز به مقایسه‌های زیاد به خاطر خاصیت دسترسی مستقیم است و مشکل بزرگ آن پیدا کردن تابع درهمی مناسب برای تولید جدول درهم به گونه‌ای که تکراری تولید نکند و سریع باشد است. البته روش‌هایی برای تولید چنین توابعی توضیح داده شده است [۵].

برنامه‌ی spell مربوط به سیستم عامل Unix از جمله برنامه‌هایی است که از یک جدول درهم برای مراجعه‌ی سریع به لغت‌نامه استفاده می‌کند[۶].

۲.۳. خلاصه تشخیص خطای غیر لغت

در کل با اینکه تحلیل چندنگاشتی ممکن از برای خطاهای تولید شده توسط ماشین، برای مثال تشخیص نوری حروف مفید باشد، ولی ثابت شده است برای خطاهای تولید شده توسط انسان دقت کم‌تری دارد. به همین خاطر بیش‌تر روش‌های تصحیح خطای املایی بر اساس مراجعه به لغت‌نامه هستند. از آنجایی که وقتی لغت‌نامه بزرگ می‌شود، سرعت پاسخ‌گویی مسئله‌ی مهمی می‌شود، روش‌هایی که از جدول‌های درهم، درخت‌ها و ... استفاده می‌کنند به وجود آمده‌اند.

۳. تصحیح خطای کلمات منزوی

از آنجایی که تشخیص تنهای خطا برای کاربردهای بسیاری کافی نیست، مسئله‌ی تصحیح خطاهای پیدا شده مطرح می‌شود. خواص کاربردهای مختلف و تفاوت آن‌ها منجر به تولید انواع متفاوتی برای کاربردهای متفاوت تصحیح کننده‌ی املایی تولید شود.

برای کاربردهای متفاوت باید سه مسئله‌ی اصلی را در نظر گرفت: (۱) مسائل مربوط به لغت‌نامه، (۲) مسائل مربوط به رابط ماشین و انسان[۵] و (۳) مسائل مربوط به الگوهای خطای املایی. مسائل مربوط به لغت‌نامه شامل طول لغت‌نامه، محدوده‌ی لغت‌نامه، نرخ ورود داده‌ی جدید به لغت‌نامه و ...

---

[۴] Hash table
[۵] Computer-human interface

مسائل مربوط به رابط ماشین و انسان شامل مسئله‌ی نیاز به پاسخ آنی، گرفتن بازخورد از کاربر، میزان دقت مورد نیاز و ...

مسائل مربوط به الگوهای خطای املایی شامل، پر تکرارترین خطاها، تعداد خطاهایی که ممکن است در یک کلمه رخ دهد، آیا خطاها طول کلمه را تغییر می‌دهند یا خیر و ...

۳.۱. الگوهای خطای املایی

الگوهای خطای املایی ممکن است بسیار با هم متفاوت باشند بسته به اینکه در چه زمینه‌ای کار می‌کنیم. برای مثال غلط‌های ناشی از استفاده از صفحه کلید (برای مثال حرف ب به جای حرف ی) بسیار متفاوت از غلط‌های ایجاد شده در تشخیص حروف نوری (برای مثال حرف لام به جای الف) است.

الگوهای خطا به سه دسته‌ی کلی تقسیم می‌شوند: (۱) خطاهای چاپی[۶]، (۲) خطاهای شناختی[۷] (۳) خطاهای آوایی[۸]. در خطاهای چاپی (برای مثال *سللم* به جای *سلام*) فرض بر این است که نویسنده یا ماشین‌نویس املای صحیح را می‌داند ولی خطا انجام می‌دهد. علت خطاهای شناختی (برای مثال *توجیح* به جای *توجیه*) عدم وجود دانش کافی در نظر گرفته می‌شود. خطاهای آوایی (برای مثال *اصلاح* به جای *اسلاح*) نوعی از خطای شناختی است، ولی کلمه‌ی خطا از لحاظ آوایی مشابه کلمه‌ی صحیح مورد نظر است.

۳.۱.۱. خطاهای املایی اصلی

در یکی از تحقیقات اولیه در زمینه‌ی خطاهای املایی نشان داد که تقریباً ۸۰ درصد از کل خطاهای املایی شامل فقط یک مورد از چهار نوع خطای اصلی املایی هستند: اضافه شدن یک حرف، پاک شدن یک حرف، جابه‌جایی دو حرف با هم، یک حرف به جای حرفی دیگر[۷].

البته تشخیص حروف نوری از این الگو پیروی نمی‌کنند. ولی مشاهدات نشان داده است که اکثر خطاها ناشی از خطای یک حرف به جای حرفی دیگر است.

۳.۱.۲. تأثیر طول کلمه

یک یافته‌ی کلی دیگر این بود که مشاهده می‌شد، طول کلمه‌ی خطا در فاصله‌ی دوتایی از طول کلمه‌ی صحیح است. این موضوع باعث شد که محققان گاهاً لغت‌نامه‌ها را بر اساس طول کلمات به چند دسته تقسیم کنند تا زمان جستجو کاهش یابد.

---

۶ Typographic
۷ Cognitive
۸ Phonetic

با توجه به قانون زیپف [۸] کلمات کوتاه تعداد تکرار بیش‌تری دارند. همچنین کلمات با تکرار بیش‌تر همسایه‌های تک خطایی بیش‌تری دارند (تصحیح کار مشکل‌تریست). ولی معمولاً کلمات کوتاه اشتباه نوشته نمی‌شوند. در ضمن هرچه کلمه کوچک‌تر باشد متن درون کلمه‌ای کمتری در اختیار ما قرار می‌دهد و کار تصحیح سخت‌تر می‌شود. مشاهده شده است که در حالی که کلمات کوتاه فقط ۹.۲ درصد از خطاها را تشکیل می‌دهند، ولی معمولاً نزدیک به ۴۲ درصد تصحیح‌های اشتباه مربوط به آن‌هاست.

۳.۱.۳. خطاهای جایگاه اول

عموماً قابل قبول است که معمولاً خطای بسیار کمتری در حرف اول کلمات رخ می‌دهد[۹]. این موضوع این فرصت را به ما می‌دهد که بتوانیم لغت‌نامه‌ها را بر اساس حرف اول تقسیم بندی کنیم تا زمان مراجعه به لغت‌نامه کاهش یابد. البته در این صورت دسته‌ای از خطاه را سیستم ما دیگر نمی‌تواند تشخیص دهد.

۳.۱.۴. تأثیر صفحه‌کلید

تحقیقات زیادی بر روی مدل کردن ماشین‌نویسی انسان انجام شده است برای مثال [۱۰]. مشاهده شد که بیش از ۵۸ درصد خطاها برای کلیدهای مجاور بر روی کیبورد بوده‌اند.

۳.۱.۵. خطاهای آوایی

نکته‌ی اصلی در این گونه خطاها بر این نکته استوار است که برای کلمات نا آشنا برای مثال اسم‌های خاص، مردم به آوای مربوط به کلمه برای املاء آن رجوع می‌کنند[۱۱]. از این لحاظ شباهت‌های آوایی حرف و هجاها (مخصوصاً در زبان فارسی) مجال مناسبی است برای انتخاب نامزدهای بهتر برای تصحیح.

۳.۲. روش‌های تصحیح خطای کلمات منزوی

مسئله‌ی تصحیح خطاهای کلمات منزوی را می‌توان به سه زیر مسئله تقسیم کرد: (۱) تشخیص خطا، (۲) تولید فهرست تصحیح‌های نامزد و (۳) رتبه‌بندی تصحیح‌های نامزد. مرحله‌ی تضخیص خطا معمولاً شامل امتحان کردن وجود تمامی چندنگاشت‌های آن یا وجود آن در یک لغت‌نامه است.

۳.۲.۱. روش‌های کم‌ترین فاصله‌ی تصحیح

بیش‌تر از هر روش دیگری در این زمینه تحقیقات انجام شده است. در [۱۲] روش‌های مختلف مقایسه‌ی رشته‌ها بیان شده است. فاصله در این مبحث به تعداد عملیات‌های تصحیح مورد نیاز گفته می‌شود. در حالت کلی در چنین روش‌هایی باید به اندازه‌ی لغت‌نامه مقایسه رشته‌ای انجام شود. البته می‌توان ابتدا یک زیر لغت‌نامه تولید کرد (لیست نامزدها)، بعد مقایسه‌ها را برای رتبه‌بندی انجام داد.

۳.۲.۲. روش‌های کلیدهای مشابه

از آنجایی که یکی از روش‌های اصلی تولید متن به صورت ماشین‌نویسی است و در ماشین‌نویسی از صفحه‌کلید استفاده می‌شود، می‌توان بر اساس صفحه کلید، و مشابهت‌های حرف کلمات را در دسته‌های مشابه قرار داد تا برای انتخاب کاندید بتوان بهتر عمل کرد. در سال ۱۹۱۸ یکی از ابتدایی‌ترین این روش‌ها ابداع شد که حرف را به صورت شکل زیر دسته‌بندی می‌کرد[۱۳].

$$\begin{aligned}
&A, E, I, O, U, H, W, Y \rightarrow 0\\
&B, F, P, V \rightarrow 1\\
&C, G, J, K, Q, S, X, Z \rightarrow 2\\
&D, T \rightarrow 3\\
&L \rightarrow 4\\
&M, N \rightarrow 5\\
&R \rightarrow 6
\end{aligned}$$

دسته‌ها بدین صورت تهیه می‌شوند: حرف اول کلمه را می‌نویسیم. برای بقیه‌ی حروف هم با توجه به جدول بالا جایگذاری می‌شوند. آنگاه صفرها حذف می‌شوند و تکراری‌های پشت سر هم یکی می‌شوند.

این روش تنها یکی از روش‌های بسیار زیاد موجود در این زمینه است.

۳.۲.۳. روش‌های آماری

روش‌های آماری از مؤثرترین این روش‌ها هستند. از اساسی‌ترین آن‌ها در [۱۵] مطرح شده است. در این روش بعد از تولید فهرست نامزدها، قصد رتبه‌بندی آن‌ها را داریم. X یکی از نامزدهای ما و Y آنچه مشاهده شده‌است است. بر اساس قانون بیز داریم:

$$P(X|Y) = \frac{P(Y|X)^* P(X)}{P(Y)}$$

P(X|Y) احتمال مناسب بودن نامزد X است که همان رتبه‌ی نامزد در نظر گرفته می‌شود. P(Y|X) احتمال وقوع خطای املایی که نتیجه‌ی آن Y است، در صورتی که منظور X بوده است را بیان می‌کند. P(X) احتمال وقوع X است که می‌توان تعداد نرمال شده‌ی آن در پیکره را در نظر گرفت.

باید برای لیست نامزدها این رتبه‌بندی انجام شود.

۴. نتیجه‌گیری

در این پژوهش تعدادی از روشها و سامانه‌هایی که برای خطایابی املایی طراحی شده‌اند، مورد بررسی قرار گرفت. برای انجام خطایابی املایی در زبان فارسی باید از میان روشهای موجود، مناسبترین روش با توجه به ویژگی‌های زبان فارسی انتخاب شود. این انتخاب از عهده‌ی این پژوهش خارج است و امید می‌رود در آینده این کار انجام شود.

۵. کارهای آینده

با توجه به محدودیت زمانی در انجام این پژوهش، فرصت بررسی روش‌های وابسته به متن و همچنین پیاده‌سازی وجود نداشت. امید است به فصل خدا در آینده فعالیت‌های مناسبی در این راستا انجام شود.

۶. منابع

[1] K. Kukich, “Techniques for automatically correcting words in text,” *ACM Computing Surveys,* vol. 24, pp. 378-439, 1992.

[2] Hanson, Riseman, and Fisher, “Context in word recognition, ” Patt. Recog. 8 35-45

[3] J. J. Hull, and S. N. Srihari, “Experiments in text recognition with binary n-gram and Viterbi algorithms,” IEEE Trans. Patt. Anal. Machine Intell. PAMI-4, 5 (Sept.), 1982 520-530.

[4] D. E. Knuth, “The Art of Programming Vol. 3, Sorting and Searching,” Addison-Wesley, Reading 1973.

[5] E. A. Fox, Q. F. Chen, and L. S. Heath, “A faster algorithm for constructing minimal perfect hash functions,” In Proceedings of the 15th annual International SIGIR Meeting. ACM, New York, 1992, 266-273.

[۶] K. Atkinson, “Gnu aspell,” In Available at http://aspell.net, 2009.

[7] F. J. Damerau, “A technique for computer detection and correction of spelling errors,” Commun ACM 7, 3 (Mar.), 1964, pp. 171-176.

[8] G. K. Zipf, “The Psycho-Biology of Language,” Houghton Mifflin, Boston, 1935.

[9] J. J. Pollock, and A. Zamora, “Collecting and characterization of spelling errors in scientific and scholarty text,” J. Amer. Soc. Inf. Sci. 34, 1, 1983, pp. 51-58.

[10] J. Grudin, “Error patterns inskilled and novice transcription typing. In Cognitive Aspects of Akilled Typewriting,” W. E. Copper, Ed. Springer-Verlag, New York, 1983.

[11] B. Van Berkel, and K. DeSmedt, “Triphone analysis, A combined method for the correction of orthographical and typographical errors,” in Proceedings of the 2nd Applied Natural Language Processing Conference (Austin, Tex., Feb.) Association for Computational Linguistics (ACL) , 1988.

[12] A. V. Aho, “Algorithms for finding patterns in strings,” in Handbook of Theroretical Computer Science, J. Van Leeuwen, Ed. Elsevier Science Publishers, 1990.

[13] M. K. Odell, and R. C. Russell, U.S Patent Number 1261167, 1918.

[14] E. M. Riseman, and A. R. Hanson, “A contextual postprocessing system for error correction using binary n-grams,” IEEE Trans. Comput. C-23, (May), 480-493, 1974.

[15] W. W. Bledsoe, and I. Browning, “Pattern recognition and reading by machine,” in Proceedings of the Eastern Joint Computer Conference, vol. 16, 225-232, 1959.