# استخراج بی‌ناظر ظرفیت فعل در زبان فارسی

محمدصادق رسولی[1]، بهروز مینایی بیدگلی[2]، هشام فیلی[3]، مریم امینیان[4]
[1] دانشکدهٔ علوم رایانه، دانشگاه کلمبیا نیویورک
[2]دانشکدهٔ مهندسی کامپیوتر، دانشگاه علم و صنعت ایران
[3]دانشکدهٔ مهندسی برق و کامپیوتر، پردیس دانشکده‌های فنی، دانشگاه تهران
[4]دانشکدهٔ علوم رایانه، دانشگاه جرج واشنگتن، واشنگتن دی. سی.

## چکیده

**ظرفیت، کلیدی‌ترین مفهوم در دستور وابستگی است. از میان مقوله‌های واژگانی گوناگون در دستور وابستگی، فعل‌ها دارای کلیدی‌ترین نقش در نحو و معنای جمله هستند. فعل مرکزیت جمله را در دستور وابستگی بر عهده داشته، معنای اصلی جمله را در درون خود نهان می‌کند. در این مقاله با بررسی روش‌های مختلف استخراج بی‌ناظر ظرفیت فعل در زبان فارسی، مسائلی درخصوص یافتن فعل در متون زبانی و ابهامات موجود در شناخت ظرفیت فعل در زبان فارسی مطرح شده، راه حل‌هایی برای آن پیشنهاد شده است. طبق بررسی‌های موجود در این مقاله، الگوریتم بیشینه‌سازی امید (EM) دارای بالاترین دقت موجود در روش‌های استخراج ساخت‌های ظرفیتی فعل در زبان فارسی است. دقت الگوریتم بیشینه‌سازی امید در این کار بیش از دو برابر آزمون فرض دوجمله‌ای بر اساس معیار F شده است.**

**واژگان کلیدی: دستور وابستگی، ظرفیت فعل، زبان فارسی، استخراج بی‌ناظر، الگوریتم بیشینه‌سازی امید (EM).**

## ۱- مقدّمه

در دستور مبتنی بر واژگان، هدف، شناخت رفتار واژه‌ها در متن از نظر نحوی یا معنایی است. در واژگان‌ها این رفتار به‌صورت دستی، خودکار یا نیمه‌خودکار درج می‌شود. شبکه‌های واژگانی، واژگان نحوی و معنایی ازجمله نمونه‌هایی برای یک واژگان مفید برای زبان هستند که باعث رشد زبان در زمینه‌های آموزشی و پردازشی می‌شوند. در این مقاله به‌جهت اهمیت ذکر شده و نیاز به دادگان واژگانی نحوی برای زبان فارسی به شناخت بی‌ناظر ظرفیت نحوی فعل در زبان فارسی پرداخته شده است. ظرفیت مفهومی است که اوّلین‌بار در نظریهٔ دستور زبان وابستگی[1] مطرح شد. نظریهٔ دستور وابستگی یکی از نظریه‌های ساخت‌گرا[2] و صورت‌گراست[3] که به طوراساسی در آن از طریق بررسی روابط وابستگی بین عناصر هسته و وابسته در زبان، به توصیف ساخت‌های نحوی در زبان‌های گوناگون پرداخته می‌شود (طبیب‌زاده، ۱۳۸۵).

با اوج گرفتن و کارآمدشدن هوش مصنوعی در آغاز دههٔ نود میلادی، کم‌کم دانش زبان‌شناسی رایانه‌ای به سمتی میل کرد که قوانین زبان‌شناسی جایشان را به روش‌های هوشمند با استفاده از روش‌های آماری و یادگیری دادند. به‌طوری‌که عمدهٔ روش‌های اخیر در زمینهٔ کشف اطلاعات زبانی، مبتنی بر روش‌های آماری بوده است (Abney, 2010). عمدهٔ این روش‌ها، از روش‌های یادگیری خودکار[4] و نظریهٔ اطلاعات[5] اقتباس شده‌اند و به‌طور دقیق همین روش‌ها در کاربردهای مختلف پردازش زبان طبیعی مورد استفاده قرار گرفته‌اند. یادگیری خودکار و پردازش زبان طبیعی هر دو شاخه‌هایی از هوش مصنوعی هستند که از جهات مختلف با هم نزدیکی بسیار دارند. درواقع می‌توان گفت که بخش

[1] Dependency Grammar Theory
[2] Structuralist
[3]Formalist

[4] Machine Learning، در عمدهٔ مراجع علمی فارسی این اصطلاح به صورت «یادگیری ماشینی» ترجمه شده است که از نظر نگارندگان اصطلاح «یادگیری خودکار» گویاتر و روان‌تر است.

[5] Information Theory

اعظمی از پردازش زبان طبیعی، یادگیری خودکار و استنتاج هوشمند بر اساس یادگیری خودکار ساخت‌های زبانی است. یادگیری خودکار با سه روش باناظر، بی‌ناظر و نیمه‌ناظر انجام می‌شود (Alpaydin, 2010).

تفاوت عمدهٔ روش‌های باناظر و بی‌ناظر در نوع پیکرهٔ مورد استفاده در یادگیری خلاصه می‌شود. در پیکرهٔ مخصوص یادگیری باناظر، همهٔ موارد دستوریِ مورد نیاز نشانه‌گذاری شده است ولی در پیکرهٔ روش بی‌ناظر هیچ نشانه‌گذاری‌ای وجود ندارد. در این میان روش‌های بی‌ناظر از چالش‌برانگیزترین روش‌های یادگیری در زبان هستند. این روش‌ها از دو جهت اساسی حائز اهمیت هستند. یک جهت این است که با استفاده از روش‌های بی‌ناظر، نیاز به نشانه‌گذاری و برچسب‌زنی دستی دادگان زبانی، به‌شدت کاهش یافته، هزینه‌ها کمتر می‌شود. جهت دیگر این است که در روش‌های بی‌ناظر یادگیری ساختارهای زبانی، شباهت بسیار زیادی با نحوهٔ یادگیری کودکان وجود دارد. با توجه به این دغدغه‌ها و کمبود دادگان مناسب زبانی در زبان فارسی، نگارندگان، یادگیری بی‌ناظر یکی از ساخت‌های زبانی را برای مقالهٔ خود انتخاب کرده‌اند. این ساخت زبانی شامل شناخت فعل و ظرفیت آن در زبان فارسی است که در ادامه به تعریف‌های مربوط به آن پرداخته می‌شود.

زبان‌شناسان مختلف روش‌های مختلفی را برای ارائهٔ دستور زبان وابستگی پیشنهاد داده‌اند که در همهٔ این دستورها این فرض پایه وجود دارد که ساختار نحوی شامل واژه‌هایی است که این واژه‌ها با روابط دودویی نامتقارن، با هم در ارتباط هستند. به این روابط، ارتباط وابستگی گفته می‌شود (Kübler et al., 2009). دو فرض اساسی در نظریهٔ دستور وابستگی وجود دارد. نخست این که هر جمله یک فعل مرکزی دارد و دوم این که بر اساس نوع و تعداد متمم‌های اجباری و اختیاری، می‌توان ساخت بنیادین جمله‌هایی را که فعل در آنها به‌کار رفته است، تعیین کرد (طبیب‌زاده، ۱۳۸۵).

## ۲- ظرفیت در دستور وابستگی

مهم‌ترین مبحث در دستور وابستگی، عبارت است از مسألهٔ ظرفیت نحوی که در آن به بحث در مورد وابسته‌های فعل، اسم و صفت پرداخته می‌شود. بر اساس این نظریه، مرکز ثقل ساختاری جمله فعل است (طبیب‌زاده، ۱۳۸۵). مفهوم ظرفیت از شیمی اقتباس شده است و عبارت است از توانایی یک عنصر در ترکیب با تعداد خاصی از اتم‌های عناصر دیگر (Tesnière, 1980).

براساس ساخت‌های ظرفیتی مختلف فعل در جمله (اختیاری‌بودن یا اجباری‌بودن هر یک از متمم‌های موجود در هر ساخت ظرفیتی)، ساخت‌های بنیادین مختلفی ایجاد می‌شود. ساخت‌های بنیادین جمله به ساخت‌هایی اطلاق می‌شود که از بسط و تعریف یا از ترکیب با هم یا از تبدیل آن‌ها به هم یا به ساخت‌های مشتق یا فرعی دیگر، یا از آمیزه‌ای از دو تا یا چند تا از روش‌های گفته‌شده، بتوان تمام جمله‌های محتمل موجود در زبان را تولید کرد. استخراج چنین ساختارهایی، از زمرهٔ مهم‌ترین و ابتدایی‌ترین وظایف تحلیل نحوی است؛ زیرا این ساخت‌های محدود و تکرارشونده، تمام ساخت‌های نحوی هر زبان را تشکیل می‌دهند (Allerton, 1982). به‌عنوان مثال، فعل «صحبت‌کردن» در زبان فارسی دارای دو ساخت ظرفیتی متفاوت است. یکی از این دو ساخت به صورت <فا،(مفح)[با]،(مفح)[از|درباره|در مورد|در خصوص|در]> است، به این معنی که فاعل به‌همراه مفعول حرف اضافه‌ای اختیاری با حرف اضافهٔ «با» و مفعول حرف اضافه‌ای اختیاری با حرف اضافهٔ «از» یا «درباره» یا «در مورد» یا «در خصوص» یا «در» است.

همان‌طور که پیداست در دستور وابستگی بی‌ترتیبی حضور اجزای زبان قابل ارائه و مدیریت بوده، نیاز به ارائهٔ ساخت‌های پیچیده وجود ندارد. درواقع برای شناخت ظرفیت فعل از روی جملات زبان، باید ساخت بنیادین جملات را شناخت و از آن به ظرفیت افعال پی برد. در ساختار زیرمقوله‌ها در دستور زایشی، اصطلاح قاب نحوی[1] وجود دارد که همان ساخت بنیادین جمله با توجه به مرکزیت فعل است.

## ۳- شناخت فعل و ظرفیت فعل فارسی

همان‌طور که اشاره شد، شناخت ظرفیت فعل در زبان فارسی حائز اهمیت بسیار زیادی برای پردازش زبان طبیعی است. علاوه‌بر این، وجود ابهام در شناخت فعل در زبان فارسی یکی از مسائلی است که هنوز بر سر آن اتفاق نظر وجود ندارد. مسألهٔ دیگری که در پردازش نحوی زبان طبیعی مشکل‌زاست؛ عدم یکسانی تعریف فعل مرکب در سطح نحو

[1] Syntactic Frame

در زبان فارسی است. حتی اگر تعریف یکسانی از فعل مرکب در زبان فارسی وجود داشته باشد، ابهام بسیاری در شناخت و استخراج افعال مرکب در زبان فارسی با استفاده از رایانه وجود خواهد داشت. زبان فارسی زاییِ بسیار بالایی در تولید افعال مرکب دارد و به‌همین دلیل به‌هیچ عنوان نمی‌توان مجموعهٔ ثابتی از افعال مرکب زبان فارسی را در نظر گرفت. با توجه به محوریت فعل در بسیاری از دیدگاه‌های زبان‌شناسی رایانه‌ای در سطح نحو و معنا، نیازی اساسی در شناخت این افعال و تهیهٔ پایگاه دانش مدونی از آنها و طرح روش‌های هوشمند برای استخراج فعل مرکزی از جملات وجود دارد.

برای شناخت ظرفیت فعل، علاوه‌بر دغدغهٔ شناخت فعل، تفکیک افزوده‌ها از متمم‌های فعل و همچنین تفکیک متمم‌های دیگر اجزای جمله از متمم فعل باعث ایجاد ابهاماتی در شناخت ظرفیت می‌شود. به عنوان مثال در جملهٔ «در تهران مبنی بر صحبت‌های رییس جمهور تصمیماتی را اتخاذ کردیم». در این‌جا حرف اضافهٔ «بر» ظرفیت صفت «مبنی» و حرف اضافهٔ «در» افزودهٔ قیدی فعل «اتخاذکردن» است. لذا وجود چنین ابهاماتی موجب می‌شود که شناخت ساخت‌های نحوی ظرفیتی جمله بیش از پیش دشوار شود. این ابهامات در جملات مرکبی که حاصل از چند فعل هستند بسیار بیشتر خواهد بود، به‌طوری که شناخت فعل مرکزی جمله خود تبدیل به دغدغه‌ای می‌شود. علاوه بر این حذف در اجزای جمله که معمولاً ناشی از هم‌پایگی و یا توصیف‌های بافتی در جملات قبلی است باعث می‌شود که استخراج ظرفیت واژگانی فعل بسیار سخت و دشوار بوده، با ابهاماتی همراه باشد. برای تفکیک بندهای متممی فعل، بندهای افزودهٔ فعل و بند صفت و اسم نیز چنین مشکلاتی وجود دارد. بندهای متممی فعل از جمله متمم‌های فعل هستند که در جایگاه‌های مختلف به جای مفعول یا فاعل می‌نشینند، درحالی‌که بندهای افزوده فقط توضیحی قیدی بر جمله می‌افزایند. بند صفت و بند اسم که از نظر صوری شباهت بسیار زیادی به بند فعل دارند نیز باعث ایجاد ابهام در شناخت درست بندهای متممی فعل می‌شوند.

## ۴ – مروری بر کارهای مشابه

در مجموع می‌توان روش‌های ارائه‌شده برای شناخت بی‌ناظر ظرفیت فعل را در سه گونه دسته‌بندی کرد (Korhonen,2002):

۱. روش‌های مبتنی بر آزمون فرض آماری[1]
۲. روش‌های مبتنی بر تخمین بیشینهٔ درست‌نمایی[2]
۳. روش‌های مبتنی بر الگوریتم بیشینه‌سازی امید (EM)[3]

روش‌های مبتنی بر آزمون فرض آماری بیشتر از بقیهٔ روش‌ها مورد آزمون قرار گرفتند. درحالی‌که روش‌های مبتنی بر تخمین بیشینهٔ درست‌نمایی، دقت بالایی در داده‌های پربسامد دارند. از سویی دیگر، روش‌های مبتنی بر بیشینه‌سازی امید، عمدتاً زمانی به کار می‌آید که اطمینان کافی از درخت‌های نحوی ارائه‌شده توسط تجزیه‌گر وجود نداشته باشد و با استفاده از روش‌های کاملاً بی‌ناظر ابهام‌زدایی می‌شود. علاوه بر این سه روش، روش‌های ترکیبی نیز وجود دارند که برای بهبود هر یک از سه روش ارائه‌شده مطرح شده‌اند. یکی از نمونه‌های روش‌های ترکیبی، استفاده از معنای واژگانی برای بهبود دقت در ابهام‌زدایی ساخت‌های زیرمقوله‌ای و ظرفیتی است.

### ۴-۱- آزمون فرض آماری

هدف از آزمون فرض این است که بین یک فرض یا خلاف آن با توجّه به مؤلّفهٔ احتمالی و داده‌های موجود تصمیمی گرفته شود. در اینجا تصمیمی که گرفته می‌شود این است که آیا ساخت زیرمقوله‌ای یا ظرفیتی خاصی مربوط به فعل مورد نظر است یا خیر (Korhonen, 2002). در روش‌های مبتنی بر آزمون فرض دو گام اصلی در نظر گرفته می‌شود: ۱) تولید یک فرض اولیه برای هر ساخت نحوی؛ و ۲) تعیین فرض نهایی برای واژگان انتخاب‌شده (Korhonen, 2002). این روش‌ها در جزئیات با همدیگر تفاوت‌هایی دارند. لذا نمی‌توان تعریف یکسانی از کم و کیف این روش‌ها ارائه داد.

در مجموع در این روش‌ها چند گام اوّلیه برای به نتیجه رسیدن وجود دارد. در مرحلهٔ نخست، دادگان موجود در پیکره برای شناخت فعل‌های درون جمله‌ها پیش‌پردازش می‌شود. بسته به زبانی که برای پردازش مورد انتخاب قرار می‌گیرد، روش‌های متفاوتی به کار می‌رود. به‌عنوان مثال در (Brent, 1993) با استفاده از روش‌های ساخت‌واژی مبتنی بر قاعده فعل‌های زبان انگلیسی مورد شناسایی قرار گرفته‌اند. در برخی دیگر از روش‌ها مانند (Manning, 1993)، این گام

[1] Hypothesis test
[2] Maximum Likelihood Estimation (MLE)
[3] Expectation-Maximization (EM)

بدین صورت در نظر گرفته نشده، تنها از یک برچسب‌زن اجزای سخن[1] برای شناسایی فعل استفاده می‌شود. در مرحلهٔ مرحلهٔ دوم، اجزای نحوی جمله با استفاده از یک تجزیه‌گر احتمالاتی و یا یک تجزیه‌گر مبتنی بر قاعده شناسایی می‌شود. در مرحلهٔ سوم، با به‌دست آمدن فراوانی هر ساخت نحوی برای هر فعل و محاسبهٔ فراوانی نسبی هر یک از آن‌ها از یکی از آزمون‌های فرض آماری استفاده می‌شود. پژوهش‌گران مختلف از آزمون‌های آماری مختلفی استفاده کرده‌اند. به‌عنوان مثال، مانینگ (Manning, 1993) از روش آزمون فرض دوجمله‌ای و برنت از آزمون فرض لگاریتم نسبت درست‌نمایی[2] (Brent, 1993) استفاده کرده است.

## ۴-۲- روش‌های بیشینهٔ درست‌نمایی

دو مرحلهٔ نخست در این روش نیز مانند روش آزمون فرض است. امّا در مرحلهٔ سوم در روش‌های بیشینهٔ درست‌نمایی پیچیدگی‌های آماری خاصی وجود ندارد. تنها از فراوانی نسبی هر ساخت ظرفیتی نسبت به فراوانی کل استفاده شده، یک آستانهٔ عددی انتخاب می‌شود. هر ساخت ظرفیتی مربوط به فعل که دارای فراوانی نسبی بیشتر یا مساوی آستانهٔ عددی داشته باشد، به‌عنوان ساخت ظرفیتی درست انتخاب می‌شود. این روش در مورد فعل‌های پرکاربرد و ساخت‌های ظرفیتی پربسامد در زبان بسیار دقیق و کاراست ولی در مورد ساخت‌های ظرفیتی کم‌کاربرد، دقّت بسیار پایینی ارائه می‌دهد (Korhonen, 2002).

Expectation Maximization Algorithm

**Input:** Initial model parameters $w^0$, training data $< \tilde{x}_1, \tilde{x}_2, \dots, \tilde{x}_{\tilde{N}} >$

**Output**: Learned parameters w

$t \leftarrow 0$

***repeat***

*E Step*:

***for*** $i = 1$ *to* $\tilde{N}$ do

$\forall y \in \Upsilon_{\tilde{x}_i}, q^t(y) \leftarrow p_{w^{(t)}}(y|\tilde{x}_i) = \frac{p_{w^{(t)}}(y,\tilde{x}_i)}{\sum_{y' \in \Upsilon_{\tilde{x}_i}} p_{w^{(t)}}(y',\tilde{x}_i)}$

***end for***

M step: $w^{(t+1)} \leftarrow$

$argmax_w \sum_{i=1}^{\tilde{N}} \sum_{y \in \Upsilon_{\tilde{x}_i}} p_w(y, \tilde{x}_i)$

$t \leftarrow t + 1$

until $w^{(t)} \approx w^{(t-1)}$

$w \leftarrow w^{(t)}$

**(شکل ۱): شبه‌برنامهٔ الگوریتم بیشینه‌سازی امید**

[1] Part of speech (POS) tagger

[2] Log likelihood ratio test

## ۴-۳- الگوریتم بیشینه‌سازی امید

در روش‌های مبتنی بر الگوریتم بیشینه‌سازی امید، همهٔ حالت‌های ممکن برای ساخت نحوی یک جمله مورد ارزیابی قرار گرفته، از روش بیشینه‌سازی امید برای تخمین بهترین مؤلفه‌های احتمالی استفاده می‌شود. در نهایت نیز با انتخاب یک مقدار آستانه یا انتخاب بهترین ساخت نحوی برای هر جمله، روش بیشینهٔ درست‌نمایی یا آزمون فرض استفاده می‌شود. از جمله کارهایی که با استفاده از بیشینه‌سازی امید انجام شده است، روش استفاده‌شده در (Carroll and Rooth, 1998) و (Dębowski, 2009) است. شبه‌برنامهٔ الگوریتم بیشینه‌سازی امید در شکل ۱ نشان داده شده است.

برای ساختاریابی ظرفیتی در الگوریتم همهٔ حالت‌های ممکن برای یک ساختار ظرفیتی برای یک جمله در نظر گرفته می‌شود و با توجه به مقادیر اولیهٔ تعیین‌شده و با داشتن همهٔ جملات و ساختارهای ممکن در جمله وزن‌های احتمالاتی به روزآوری می‌شود. برای پیدا کردن حالات ممکن به طور معمول از ماتریس باهم‌آیی ساخت‌های نحوی (مانند حرف اضافه‌ها) استفاده می‌شود (Dębowski, 2009).

# ۵ – شناخت بی‌ناظر ظرفیت فعل

همان‌طور که اشاره شد، به‌جز الگوریتم بیشینه‌سازی امید در همهٔ روش‌ها با استفاده از یک تجزیه‌گر مطمئن و دارای دقت بالا از آزمون‌های فرض آماری یا استفاده از یک مقدار آستانه برای الگوریتم تخمین بیشینهٔ درست‌نمایی استفاده می‌شود. درحالی که چنین تجزیه‌گری در زمان انجام این پژوهش، برای زبان فارسی وجود نداشته است[3]. تنها منبعی که می‌توان از آن استفاده کرد، پیکرهٔ *بی‌جن‌خان* (بی‌جن‌خان، ۱۳۸۳) است که این پیکره در سطح صرف برچسب خورده است و می‌توان اطلاعاتی مانند حرف اضافه‌ها، اسم‌ها و دیگر اجزای جمله را استخراج کرد. درنتیجه ما به الگوریتمی نیاز داریم که:

۱) به جای یک حالت خروجی ساخت نحوی برای هر جمله، همهٔ حالات ممکن را بتواند در خود نگاه دارد؛

۲) ذاتاً یک الگوریتم بی‌ناظر باشد؛

[3] هم‌اکنون پیکرهٔ وابستگی زبان فارسی (Rasooli et al., 2013)، با دارا بودن حدود سی هزار جملهٔ برچسب‌خورده در سطح نحو و برچسب اجزای سخن، داده‌ای قابل اتکا برای توسعهٔ تجزیه‌گرهای نحوی است.

۳) قابلیت ارائهٔ احتمالات هر ساخت ظرفیتی را برای هر فعل داشته باشد که با آن بتوان از تخمین بیشینهٔ درست‌نمایی استفاده کرد.

بر اساس فرض‌های مطرح‌شده از الگوریتم بیشینه‌سازی امید برای استخراج ساخت‌های ظرفیتی استفاده شده است.

## ۵-۱- الگوریتم بیشینه‌سازی امید در استخراج ظرفیت فعل در زبان فارسی

در الگوریتم بیشینه‌سازی امید بین دو مرحله گردش تکرار انجام می‌شود. در یک مرحله توزیع پسینی متغیرهای پنهان $Y_i$ برای داده‌های مشاهده‌شدهٔ $x_i$ محاسبه شده، در مرحلهٔ بعد وزن احتمالاتی مؤلفه‌های الگوی احتمالاتی ($w_i$) به روزرسانی می‌شود.

احتمال پسینی[1] برای دادهٔ iام برابر با $p_w(Y_i, x_i)$ خواهد بود. مرحلهٔ اول مرحلهٔ E (امیدیابی) و مرحلهٔ دوم مرحلهٔ M (بیشینه‌سازی) است (Smith, 2011). شایان ذکر است که مبنای این الگوریتم نیز دیدگاه آماری بیشینهٔ درست‌نمایی است. بدین معنا که بر اساس مشاهدات مورد نظر، بهترین احتمال ممکن به‌صورت قطعی به دست می‌آید. این الگوریتم به صورت کاهش شیب عمل می‌کند. با انتخاب یک نقطهٔ اولیه در فضای جستجو و با به‌روزشدن مقادیر مؤلفه‌های احتمالاتی به‌صورت تکرار گردشی، یک نقطهٔ بهینهٔ محلی پیدا می‌شود. یکی از ضعف‌های موجود در الگوریتم بیشینه‌سازی امید همین است که هیچ وقت ضمانتی برای پیدا شدن بهینهٔ محلی وجود ندارد و با مقادیر اولیهٔ متفاوت نتایج متفاوتی به‌دست می‌آید.

در این مقاله، مبنا بر این مسأله بوده است که با توجّه به این که ساخت‌های ظرفیتی انتزاعی بوده، درجمله نمود پیدا نمی‌کنند و آن چه در جمله نمود پیدا می‌کند ساخت بنیادین جمله است (طبیب‌زاده، ۱۳۸۵)، ساخت‌های بنیادین ممکن برای فعل اوّل جمله استخراج شود. دلیل این که فعل اوّل جمله مورد بررسی قرار گرفته است، این بوده که در جملات چندفعلی، به‌طور معمول فعلی که بند متممی یا موصولی است، بعد از فعل اصلی می‌آید؛ مانند جملهٔ «او گفت که می‌آید». برای ساده‌سازی و کاهش پیچیدگی، ساخت‌های هم‌پایگی به‌عنوان وجه ابهام مسأله در نظر گرفته شده، از بررسی آنها به‌صورت مستقل پرهیز شده است. بدین ترتیب برای هر زوج ساخت ظرفیتی و فعل احتمال توأم مورد بررسی و تخمین در نظر گرفته شده است. در رابطهٔ (۱)، این مقدار احتمالاتی نمایش داده شده است که در آن q تابع فراوانی احتمالاتی است. val نشان‌دهندهٔ یک ساخت بنیادین و verb نشان‌دهندهٔ فعل خاص است.

$$P(val_i, verb_j) = \frac{q(val_i, verb_j)}{q(verb_j)} \quad (۱)$$

در الگوریتم بیشینه‌سازی امید در مرحلهٔ اوّل برای هر فعل و ساخت ظرفیتی یک مقدار اوّلیه در نظر گرفته می‌شود و سپس با استفاده از گردش تکرار در مراحل امیدیابی و بیشینه‌سازی این مقادیر احتمالاتی به‌روز می‌شود. درنتیجه برای هر فعل رابطهٔ (۲) صدق می‌کند.

$$\sum_i P(val_i, verb_j) = 1 \quad (۲)$$

در زمانی که تعداد تکرارهای الگوریتم به‌حدّی برسد که احتمالات هم‌گرا شوند، با استفاده از یک مقدار آستانه که به‌صورت دستی تنظیم می‌شود، ساخت‌های ظرفیتی ممکن برای هر فعل انتخاب شده، نتایج مورد ارزیابی قرار می‌گیرند.

## ۶ – آزمایش و ارزیابی

پس از فعل‌یابی با ابزار متن‌باز فعل‌یاب زبان فارسی (Rasooli et al., 2011a)، برای هر جمله با استفاده از تجزیه‌گر مبتنی بر سرنخ‌های ساده، ساخت‌های بنیادین ممکن هر جمله به‌دست می‌آید. درنتیجه فراوانی احتمالاتی هر ساخت بنیادین بر اساس ساخت‌های بنیادین دیگر به روزآوری می‌شود. به عنوان مثال در جملهٔ «علی را در پارک دیدم»، با توجه به وجود «را» و «در» در جمله، دو ساخت بنیادین ||فا، مف|| و ||فا، مف، مفح[در]||[2] دیده می‌شود، اگر احتمال ساخت بنیادین ||فا، مف|| برای فعل دیدن ۰/۲۵ و احتمال ساخت بنیادین ||فا، مف، مفح[در]|| ۰/۱۵ باشد، آن‌گاه فراوانی احتمالی ساخت اول برای جملهٔ مذکور $\frac{۲۵}{۴۰}$ می‌شود. به بیانی دیگر، برای هر جمله با توجه به تعداد ساخت‌های بنیادین ممکن فراوانی احتمالاتی هنجارسازی می‌شود. با توجه به این که در این تجزیه‌گر توان تمیز بین دو ساخت بنیادین ممکن وجود ندارد، استفاده از الگوریتم

[1] Posterior probability

[2] «فا» نشان‌دهندهٔ فاعل، «مفح» نشان‌دهندهٔ مفعول حرف اضافه‌ای و «مف» نشان‌دهندهٔ مفعول است.

بیشینه‌سازی امید، شاید بهترین گزینه در بین الگوریتم‌های یادشده در بخش چهارم باشد. در الگوریتم بیشینه‌سازی امید فراوانی‌های احتمالاتی محاسبه می‌شود و بدین صورت این نیاز که بتوان برای یک جمله به صورت همزمان چند ساخت بنیادین مختلف را در نظر داشت، برآورده خواهد شد.

موارد زیر برای انتخاب ساخت‌های بنیادین ممکن در نظر گرفته شده است:

- در اوّلین مرحله از الگوریتم بیشینه‌سازی امید، احتمال هنجارنشدهٔ هر ساخت بنیادین در هر جمله، نسبت مستقیم با تعداد خوشه‌های نحوی موجود در ساخت بنیادین دارد. به عنوان مثال در ساخت‌هایی که چهار خوشهٔ نحوی (فا، مفح۱، مفح۲، مف) وجود دارد، احتمال چهار برابر حالتی است که تنها یک خوشهٔ نحوی در ساخت وجود دارد (فا).
- اگر وجود «را» یا ساخت مجهول در جمله ثابت شد، احتمال ساخت‌های بی‌مفعول در آن جمله صفر خواهد شد.
- اگر حرف اضافه در جمله وجود داشته باشد، احتمال وجود ساخت بنیادین بدون حرف اضافه کاهش می‌یابد (در اینجا به‌صورت دستی تقسیم بر ۱۰ شده است).

### ۶-۱- تهیهٔ دادهٔ معیار برای ارزیابی

دادهٔ معیار برای ارزیابی نتایج این پژوهش، فرهنگ ظرفیت فعل در زبان فارسی (Rasooli et al., 2011b) است که نسخهٔ اوّل آن شامل ۴۲۸۲ فعل و ۵۴۲۹ زوج ساخت ظرفیتی و فعل منحصر به فرد است.

### ۶-۲- پیش پردازش دادهٔ آموزشی و برچسب‌زنی اجزای سخن

در این پژوهش برای تهیهٔ دادهٔ یادگیری از پیکرهٔ بی‌جن‌خان (بی‌جن‌خان، ۱۳۸۳) استفاده کردیم. به‌عنوان پیش‌پردازش نخست ساخت‌های بنیادین موجود در ساخت‌های ظرفیتی فرهنگ ظرفیت استخراج شده‌اند. در بین ساخت‌های بنیادین ممکن در فرهنگ ظرفیت، تنها ساخت‌های بنیادینی که حداقل یک بار درون پیکره به‌عنوان یک ساخت بنیادین ممکن دیده شده‌اند، به رسمیت شناخته شده است.

ساخت‌های نحوی با متمم قیدی را به‌دلیل معنایی‌بودن و ساخت‌های تمییزی و مسندی را به‌دلیل کم‌کاربرد و محدودبودن در نظر نگرفته‌ایم. ساخت‌های متمم قیدی در زمانی مانند جملهٔ «تهران رفتم» به جای «به تهران رفتم» که تشخیص این که تهران نام یک مکان است فارغ از تشخیص ساخت‌های نحوی است.

چنین پدیده‌ای در زبان انگلیسی بدین صورت وجود ندارد و این مسأله یکی از وجه‌های ابهام‌زای زبان فارسی است. از آن جایی که در پیکرهٔ بی‌جن‌خان کلمات بسیار زیادی به عنوان حرف اضافه در نظر گرفته شده‌اند، تنها حرف اضافه‌های پرکاربرد زبان فارسی مانند «از»، «به»، «در»، «تا» و غیره به‌عنوان گزینه‌های ممکن برای حرف اضافه در نظر گرفته شده‌اند. با وجودی که در نظر گرفتن برخی از حرف اضافه‌ها باعث ایجاد خدشه در یادگیری می‌شود، ولی وجود چنین حرف اضافه‌های کم‌کاربردی به افزایش دقت استخراج‌گر کمک شایانی نخواهد کرد و حتی ممکن است باعث افت کارایی آن شود. در مرحلهٔ پایانی نیز با استفاده از ابزار فعل‌یاب (Rasooli et al., 2011a)، پیکرهٔ بیجن‌خان پیش‌پردازش شده است. در این صورت به‌عنوان مثال دنبالهٔ واژه‌هایی مانند «گفته شده است»، «خواهم گفت»، «گفته خواهد شد»، «می‌گویم» و «گفته‌اند» همه به ریشهٔ‌شان به‌صورت «گفت#گو» تبدیل می‌شود. بدین صورت با ریشه‌یابی فعل فضای احتمالاتی داده پالایش می‌شود.

## ۷ – نتایج

در این آزمایش الگوریتم بیشینه‌سازی امید با مقادیر آستانهٔ مختلف مورد آزمون قرار گرفته است. مقدار احتمال در نظر گرفته شده برای هر یک از ساخت‌های ظرفیتی نیز برابر در نظر گرفته شده است. دادهٔ آزمون شامل فعل‌هایی می‌شدند که حداقل دویست بار در پیکره دیده شده باشند. از این میان ساخت‌هایی که به‌هیچ‌وجه در پیکره دیده نشدند، مورد ارزیابی قرار نگرفتند.

در این آزمایش از یکی از روش‌های ابتکاری موجود برای فرار از بیشینهٔ محلی نیز در الگوریتم بیشینه‌سازی امید استفاده شده است که شامل بازتعریف تصادفی[1] برخی از مقادیر مؤلّفه‌های احتمالی است. این مفهوم از فلسفهٔ شبه‌تپه‌نوردی الگوریتم بیشینه‌سازی امید نشأت می‌گیرد.

[1] Random resetting

در واقع با بازتعریف برخی از مقادیر احتمالات به صورت اتفاقی تپه‌نورد مکان خود را کاملاً اتفاقی تغییر می‌دهد تا خطر افتادن سریع در بیشنیهٔ محلی از بین برود.

در این روش بر اساس شمارهٔ گردش تکرار الگوریتم و یک عدد تصادفی، برای هر خوشه به احتمال متناسب با عدد تصادفی ممکن است مقدار احتمال بازتعریف شود. بدین صورت نتایج ممکن است جابه‌جا شود. در شکل ۲، شبه‌برنامهٔ این روش نشان داده شده است.

Random Resetting Algorithm for Valence Induction

*for each possible valency slot in sentence*

$$rand\text{-}prob=\frac{1}{\text{iteration-number}}$$

*" if(rand-number≤rand-prob')*

$$p(x_i,y_i)=a\ new\ random\text{-}generated\text{-}number$$

*end*

*end-for*

**(شکل ۲): شبه‌برنامهٔ الگوریتم بازتعریف تصادفی**

در جدول ۱، نتایج آزمایش با مقادیر آستانهٔ مختلف نشان داده شده است. بر اساس معیار F، مقدار ۰/۱ به عنوان حد آستانهٔ قابل قبول انتخاب شد.

برای این که مقایسه‌ای بین روش بیشینه‌سازی امید، بازتعریف تصادفی و روش‌های اصلی داشته باشیم، آزمون فرض دو جمله‌ای را مورد آزمون قرار دادیم. نتایج حاصل در جدول ۲ نمایش داده شده است. در این جدول از حد آستانهٔ ۰/۱ استفاده شده است. همان‌طور که در این جدول نشان داده شده‌است، در آزمون فرض هیچ تصمیم درستی در مورد صحت یک ساخت ظرفیتی در یک فعل گرفته نشده، همهٔ ساخت‌های ممکن به‌عنوان ساخت‌های محکم و قابل اتکا برای فعل در نظر گرفته شده‌است.

**(جدول ۱): نتایج به‌دست آمده از کارایی الگوریتم بیشینه‌سازی امید در استخراج ظرفیت فعل فارسی**

| مقدار آستانه | دقت | فراخوانی | سنجهٔ F |
|---|---|---|---|
| **۰/۰۵** | ۴۸ | ۴۲/۴۷ | ۴۵/۰۷ |
| **۰/۱** | ۶۱/۱۱ | ۳۸/۹۳ | ۴۷/۵۶ |
| **۰/۱۵** | ۶۵/۰۷ | ۳۶/۲۸ | ۴۶/۵۹ |
| **۰/۲** | ۶۲/۷۴ | ۳۱/۲۸ | ۳۹/۰۲ |
| **۰/۲۵** | ۶۶/۶۷ | ۲۴/۷۷ | ۳۶/۱۲ |
| **۰/۳** | ۶۶/۶۷ | ۲۴/۷۷ | ۳۶/۱۲ |
| **۰/۳۵** | ۷۱/۷۹ | ۲۴/۷۷ | ۳۶/۸۴ |
| **۰/۴** | ۶۹/۴۴ | ۲۲/۱۲ | ۳۳/۵۵ |
| **۰/۴۵** | ۷۰/۵۸ | ۲۱/۲۳ | ۳۲/۶۵ |
| **۰/۵** | ۷۰ | ۱۸/۵۸ | ۲۹/۳۷ |

دلیل اصلی این نااستواری در نتایج، ضعف در استخراج ساخت پیشنهادی از سوی تجزیه‌گر بوده است. همین مسأله دلیلی بر این است تا زمانی یک تجزیه‌گر مناسب برای جملات زبان وجود نداشته باشد، استخراج ساخت‌های ظرفیتی به‌صورت مستقل کاری بسیار دشوار خواهد بود. نکتهٔ دیگر در این مورد، دشواری تشخیص ساخت‌های متممی از ساخت‌های افزوده‌ای است. در ساخت‌های افزوده‌ای، هیچ تفاوت صوری‌ای از نظر جای‌گیری اجزای نحوی زبان وجود نداشته، تنها تفاوت در معنای متممی اجزای متممی زبان است درحالی‌که در اجزای افزوده‌ای با حذف هر یک افزوده‌ها خللی به خوش‌ساختی جمله وارد نمی‌شود. مسألهٔ بعدی اختیاری‌بودن متمم‌ها است. با اختیاری‌بودن متمم‌ها، امکان تشخیص صحت یک ساخت برای تشخیص‌دهنده، سخت‌تر نیز می‌شود.

**(جدول ۲): کارآیی روش‌های مختلف در استخراج ساخت‌های ظرفیتی**

| | دقت | فراخوانی | سنجهٔ F |
|---|---|---|---|
| **آزمون فرض دوجمله‌ای** | ۱۳/۰۲ | ۸۲/۲۴ | ۲۲/۴۷ |
| **بیشینه‌سازی امید** | ۶۱/۱۱ | ۳۸/۹۳ | **۴۷/۵۶** |
| **بازتعریف تصادفی** | ۶۰/۸۷ | ۳۸/۱۸ | ۴۶/۹۷ |

## ۷-۱- مقایسه با دیگر زبان‌ها

دقّت موجود در سایر زبان‌ها به‌طور کامل وابسته به نوع زبان، فعل‌های در نظر گرفته‌شده و پیکرهٔ مورد آزمون است. در زبان‌هایی مانند انگلیسی تنها از ساخت‌هایی که حداقل دویست بار در پیکره دیده شده‌اند برای آزمون روش استفاده شده است. در جدول ۳، کارایی روش‌های مختلف گزارش شده‌است. شایان ذکر است که این کارایی‌ها با توجه به این بوده است که در زبان انگلیسی تجزیه‌گر احتمالاتی نحوی وجود دارد و گوناگونی تصریف در زبان انگلیسی مانند زبان فارسی نیست.

(جدول ۳): کارایی روش‌های اجراشده برای یافتن ظرفیت افعال انگلیسی [۲۸]

| | دقت | فراخوانی | معیار F |
|---|---|---|---|
| آزمون فرض دوجمله‌ای | ۵۰/۳ | ۵۶/۶ | ۵۳/۳ |
| آزمون فرض نسبت لگاریتمی درست‌نمایی | ۴۲/۳ | ۴۸/۴ | ۴۵/۱ |
| بیشینهٔ درست‌نمایی | ۷۴/۸ | ۵۷/۸ | ۶۵/۲ |

## ۷-۲- تحلیل نتایج

یکی از نقاط ضعف موجود در نتایج، عدم شناخت ساخت‌های مفعول حرف اضافه‌ای است. دلیل اصلی در عدم شناخت این ساخت، کم‌بسامدبودن این ساخت به‌صورت خاص برای هر فعل است. بدین صورت ساخت‌های دارای دو مفعول حرف اضافه‌ای مورد شناسایی قرار نگرفته، دارای اتکای آماری قابل قبول نبوده‌اند. از سویی دیگر ضعف ناشی از پیچیدگی زبان در تجزیه‌گر نحوی مبتنی بر قاعده باعث شده است که دقت استخراج ساخت‌های ظرفیتی تفاوت ویژه‌ای با استخراج به‌طور کامل تصادفی ساخت‌ها نداشته باشد.

در الگوریتم بیشینه‌سازی امید، با در نظر داشتن این مهم، همهٔ ساخت‌های بنیادین ممکن برای جمله به رسمیت شناخته شده، در یک مرحله امید هر ساخت بنیادین استخراج شده، در مرحله‌ای دیگر این مقادیر احتمالاتی با توجه به داده‌های مشاهده‌شده بیشینه می‌شود. درواقع در روش آزمون فرض این امکان که به هر خروجی از تجزیه‌گر یک وزن تعلق داد وجود نداشته، همهٔ حالات ممکن تجزیه دارای یک وزن برابر هستند درحالی‌که در الگوریتم بیشینه‌سازی امید این مسأله با احتمالاتی‌کردن فراوانی‌ها مرتفع شده است.

حسن روش مبتنی بر الگوریتم بیشینه‌سازی امید مبتنی بودن این روش‌ها بر عدم قطعیت موجود در مسأله است. البته در این الگوریتم نیز در شناخت ساخت‌های کم‌بسامد (مانند ساخت‌های با دو مفعول حرف اضافه‌ای) ضعف وجود دارد.

## ۸- جمع‌بندی

استخراج ساخت‌های زبانی برای زبان فارسی با توجّه به بی‌ترتیبی و پرتصریفی بسیار پیچیده‌تر از زبان‌هایی مانند انگلیسی است. طیف وسیع روش‌های پیشنهادی مطرح‌شده در این پژوهش حاکی از این مسأله است که در زمینهٔ پردازش زبان فارسی در سطح دستور واژگانی و دستور وابستگی، کارهای زیادی انجام نگرفته و نیاز است پژوهش‌گران به این دستور توجه بیشتری نمایند.

طیف وسیع سبک نگارش کلمات در زبان فارسی، سبب ایجاد نوعی ناهمگونی نوشتاری در زبان می‌شود (به‌عنوان مثال «هم‌بازی»،«همبازی» و «هم بازی»؛ یا «واژه‌ی»، «واژه ی» و «واژهٔ»). این مسأله شاید از دیدگاه انسانی آن‌چنان چالش‌برانگیز نباشد، ولی از دیدگاه رایانه‌ای موجب ایجاد خدشه در نظم داده‌ها می‌شود و تعمیم‌پذیری الگوریتم‌های یادگیرنده را زیر سؤال می‌برد. از نظر نگارندگان، اگر این مسأله به‌صورت پیش‌پردازشی با قابلیت اتکای بالایی قابل حل شود، سطح پردازشی زبان فارسی رشد و شکوفایی بسیار بالایی خواهد داشت.

ضعف موجود در دیدگاه‌های زبان‌شناسیِ مخصوص زبان فارسی، باعث شده است که از نظر پردازشی نیز با ضعف‌هایی مواجه باشیم. نیاز به بسترهای دادگانی مانند شبکه‌های فعلی از قبیل وربنت (*Kipper-Schuler*, 2005)، شبکه‌های واژگانی، پیکره‌های معنایی و گفتمانی در زبان فارسی بیش از پیش احساس می‌شود. به بیانی دیگر، اگر بتوان بسترهای محتوایی قابل آزمون برای زبان فارسی بیش از چیزی که هم‌اکنون در دسترس است، فراهم آورد و از روش‌های دستی و مبتنی بر قاعده‌های سلیقه‌ای پرهیز کرد، شاهد پیشرفت‌های چشم‌گیری در پردازش زبان فارسی خواهیم بود.

## ۹- مراجع


بیجن‌خان، م.، نقش پیکره‌های زبانی در نوشتن دستور زبان: معرفی یک نرم‌افزار رایانه‌ای. مجلهٔ زبان‌شناسی ایران، سال ۱۹، شمارهٔ ۲: صص. ۴۸-۶۷، مرکز نشر دانشگاهی، تهران، ایران.

طبیب‌زاده، ا.، ظرفیت فعل و ساخت‌های بنیادین جمله در فارسی امروز. ۱۳۸۵: نشر مرکز.

Abney, S., "Statistical Methods in Language Processing". Wiley Interdisciplinary Reviews: Cognitive Science 2010.

Allerton, D.J., Valency and the English Verb. 1982, London: Academic Press.

Alpaydin, E., "Introduction to machine learning". 2nd ed. 2010: The MIT Press.

Brent, M.R., "From Grammar to Lexicon: Unsupervised Learning of Lexical Syntax". Computational Linguistics, 1993. 19(2): p. 243-262.

Carroll, G. and M. Rooth, Valence Induction with a Head-Lexicalized PCFG, in Workshop of Empirical Methods in NLP. 1998: Granada.

Dębowski, Ł., Valence extraction using EM selection and co-occurrence matrices. Lang Resources & Evaluation, 2009. 43: p. 301-327.

Korhonen, A., "Subcategorization acquisition". 2002, PhD. thesis, University of Cambridge.

Kübler, S., R. McDonald, and J. Nivre, "Dependency Parsing". Syntehsis Lectures on Human Language Technologies, ed. G. Hirst. 2009: Morgan & Claypool Publishers.

Manning, C.D. "Automatic acquisition of a large subca-tegorization dictionary from corpora". 1993: Association for Computational Linguistics.

Rasooli, M. S., H. Faili, and B. Minaei-Bidgoli, "Unsupe-rvised Identification of Persian Compound Verbs". Advances in Artificial Intelligence, 2011: p. 394-406.

Rasooli, M. S., A. Moloodi, M. Kouhestani, and B. Minaei-Bidgoli, "A Syntactic Valency Lexicon for Persian Verbs: The First Steps towards Persian Dependency Treebank". 5th Language & Technology Conference (LTC): Human Language Technologies as a Challenge for Computer Science and Linguistics. 2011: Poznań, Poland. p. 227-231.

Rasooli, M. S., Kouhestani M., and Moloodi A., "Development of a Persian Syntactic Dependency Treebank", The 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL HLT), Atlanta, USA, June 2013, p. 306-314.

Kipper-Schuler, K. "VerbNet: A broad-coverage, comp-rehensive verb lexicon". PhD Thesis, University of Pennsylvania, 2005.

Smith, N.A., "Linguistic Structure Prediction". Synthesis Lectures on Human Language Techn
ologies, 2011. 4(2): p. 1-274, Morgan & Claypool Publishers.

Tesnière, L., "Grundgzüge der Strukturalen Syntax", ed. H.V.U. Engel. 1980, Stuttgart: Klettcotta.

**محمدصادق رسولی** دانشجوی دکترای علوم رایانه در دانشگاه کلمبیا در شهر نیویورک است. وی تحصیلات خود را در مقطع کارشناسی مهندسی نرم‌افزار در دانشکدهٔ مهندسی کامپیوتر دانشگاه علم و صنعت ایران در سال ۱۳۸۸ به‌پایان رساند و سپس در همان دانشکده، کارشناسی ارشد هوش مصنوعی و رباتیک را در سال ۱۳۹۰ اخذ نمود. زمینهٔ تحقیقاتی اصلی ایشان پردازش زبان طبیعی و زبان‌شناسی رایانه‌ای است.

نشانی رایانامهٔ ایشان عبارت است از:

**rasooli@cs.columbia.edu**

**بهروز مینایی بیدگلی** دکترای خود را در رشته‌ی علوم و مهندسی کامپیوتر از دانشگاه ایالتی میشیگان آمریکا در سال ۱۳۸۴ گرفت. تخصص او هوش مصنوعی و داده کاوی است و هم اکنون به‌عنوان عضو هیئت علمی دانشکدهٔ مهندسی کامپیوتر دانشگاه علم و صنعت به تدریس دروس هوش مصنوعی و نرم‌افزار مشغول است. ایشان سرپرستی گروه متن‌کاوی برای متون عربی و فارسی را در پژوهشکدهٔ متن کاوی نور نیز به‌عهده دارد. از سال ۱۳۸۶ ریاست بنیاد ملی بازی‌های رایانه‌ای بر عهدهٔ ایشان است.

نشانی رایانامهٔ ایشان عبارت است از:

**b_minaei@iust.ac.ir**

**هشام فیلی** تحصیلات خود را در مقطع کارشناسی مهندسی نرم‌افزار در دانشکدهٔ مهندسی کامپیوتر دانشگاه صنعتی شریف با رتبهٔ یک در سال ۱۳۷۶ به پایان رساند. سپس مقاطع کارشناسی ارشد نرم‌افزار و دکتری هوش مصنوعی را به‌ترتیب در سال‌های ۱۳۷۸ و ۱۳۸۵ در همان دانشکده تکمیل کرد. از سال ۱۳۸۷ تاکنون عضو هیئت علمی دانشکدهٔ مهندسی برق و

کامپیوتر دانشکدهٔ فنی دانشگاه تهران است. زمینه‌های تحقیقاتی مورد علاقهٔ ایشان پردازش هوشمند متن و زبان طبیعی، ترجمهٔ ماشینی، داده‌کاوی، بازیابی اطلاعات و شبکه‌های اجتماعی هستند.

نشانی رایانامهٔ ایشان عبارت است از:
**hfaili@ut.ac.ir**

**مریم امینیان** دانشجوی دکتری علوم رایانه در دانشگاه جرج واشنگتن در شهر واشنگتن دی. سی. است. وی تحصیلات خود را در مقطع کارشناسی نرم‌افزار در دانشکدهٔ مهندسی کامپیوتر دانشگاه علم و صنعت ایران با رتبهٔ یک در سال ۱۳۸۹ به پایان رساند. سپس مقطع کارشناسی ارشد هوش مصنوعی و رباتیک را در دانشکدهٔ مهندسی کامپیوتر دانشگاه صنعتی شریف در سال ۱۳۹۱ تکمیل کرد. زمینهٔ تحقیقاتی اصلی ایشان پردازش زبان طبیعی و زبان‌شناسی رایانه‌ای است.

نشانی رایانامهٔ ایشان عبارت است از:
**aminian@gwu.edu**