بسم الله الرحمن الرحیم

پیکرهٔ متنی بر اساس دستور وابستگی، راه‌کارها و چالش‌ها

پژوهش و گردآوری:

محمّد صادق رسولی

rasooli.ms@gmail.com

مرکز تحقیقات کامپیوتری علوم اسلامی، معاونت تهران

فهرست مطالب

عنوان شماره صفحه

۱. مقدمه

امروزه یکی از نیازهای اصلی در زمینهٔ پردازش هوشمند متون زبانی، تحلیل و فهم محتوای متنی است. رویکردهایی که در زمینهٔ پردازش رایانه‌ای نحو و معنای زبان بهترین پاسخ‌ها را دریافت کرده است، رویکردهای مبتنی بر استفاده از هوش مصنوعی و پردازش بر اساس روش‌های آماری از روی دادگان آمادهٔ زبانی بوده است [۱]. در این صورت روند کار طوری می‌شود که باید برنامه‌ای را توسعه داد که این برنامه قابلیت یاد گرفتن و فهم الگوهای زبانی را از روی پیکره‌های متنی داشته باشد.

هم‌اکنون چنین رویکردهایی در زبان‌هایی چون انگلیسی، فرانسوی، سوئدی، آلمانی و دانمارکی به صورت فعال در حال پیگیری است به طوری به تازگی مرکز پروژه‌های پژوهشی دفاعی ایالات متحده کتابی [۲] در زمینهٔ اکتشافات جدید در مورد پردازش زبان طبیعی و ترجمهٔ خودکار منتشر کرده است که طبق نوشته‌های این کتاب، یکی از دغدغه‌های اصلی، تهیهٔ دادگان مناسب جهت یادگیری الگوهای زبانی است.

۲. وضعیت دادگان زبانی در زبان فارسی

هم‌اکنون در زبان فارسی ضعف از ناحیهٔ کمبود دادگان به شدت احساس می‌شود. تاکنون تنها پیکرهٔ برچسب‌خوردهٔ مناسب برای فهم زبان، پیکرهٔ متنی بیجن‌خان است [۳] که تنها در آن اطلاعات صرفی و ساخت‌واژی زبان برچسب خورده است و اطلاعاتی در مورد ساختار نحوی و معنایی جملات وجود ندارد. به همین دلیل نیاز اساسی برای تهیهٔ دادگان نحو و معنای فارسی به چشم می‌خورد. علاوه بر این، دغدغهٔ دیگری که در زمینهٔ زبان فارسی به چشم می‌خورد، انتخاب بازنمایی مناسبی از نحو و معنا است که در آن بتوان ویژگی‌های خاص زبان فارسی مانند بی‌ترتیبی را در آن گنجاند.

۳. تهیهٔ دادگان نحوی زبان فارسی

پس از مطالعات فراوان در مورد نیازمندی امروز زبان فارسی برای فهم محتوا و معنا، به این نتیجه رسیدیم که اولین گام برای فهم متن پس از شناخت واژه، شناخت نحو جمله است. این شناخت به خودی خود یاری‌گر رایانه در شناخت مفهوم بوده، علاوه بر آن به عنوان پیش‌پردازشی برای تجزیهٔ معنایی زبان نیز تلقی می‌شود.

انتخاب نوع بازنمایی برای تهیهٔ پیکرهٔ متنی بسیار حائز اهمیت است. به عنوان مثال دو مرحله از تهیهٔ پیکرهٔ درختی چینی در مجموع ۵ سال به طول انجامید [۴]؛ لذا باید قبل از هر عملی دستور و بازنمایی مناسبی را برای پیکرهٔ زبانی انتخاب کرد. نتیجهٔ مطالعات در زمینهٔ پیکره‌های موجود با بازنمایی نحوی،

این بوده که تاکنون دو نوع بازنمایی بیش‌تر مورد توسعه قرار گرفته است. بازنمایی اول، پیکره‌های بر مبنای دستور زایشی[۱] است که معروف‌ترین پیکرهٔ موجود در این مورد، پیکرهٔ دادگان درختی پن[۲] [۵] است. در این نوع از نمایش، جمله به عبارت‌هایی و عبارات به زیرعبارات و زیرعبارت‌ها به واژه‌ها تقسیم‌بندی می‌شوند. در واقع در این گونه از نمایش جمله به سازه‌هایی[۳] تقسیم می‌شود. این نوع نمایش قابل ارتقا به سطح معنا نیز هست [۶] که به عنوان نمونه می‌توان به دادگان گزاره‌ها[۴] [۷] اشاره کرد. یکی از ضعف‌های اصلی در نمایش زایشی، عدم توانایی در نمایش صریح بی‌ترتیبی در زبان است.

بازنمایی دوم بر اساس دستور وابستگی است که از جمله پیکره‌های موجود در این دستور که هم در سطح صرف و ساخت‌واژه، هم در سطح نحو و هم در سطح معنا برچسب‌گذاری شده است، پیکرهٔ وابستگی پراگ [۸] برای زبان چکی است. این دستور به نمایش زبان انسانی نزدیک‌تر بوده، قابلیت نمایش بی‌ترتیبی زبان در آن وجود دارد [۴, ۹]. به دلیل وجود بی‌ترتیبی در زبان‌هایی مانند ترکی [۱۰]، چکی [۸]، آلمانی [۱۱]، دانمارکی [۱۲]، عربی [۱۳] و لاتین [۱۴]، دستور وابستگی برای ساخت دادگان درختی ارجحیت داده شده‌اند. در بخش بعدی این مستند، به صورت مبسوط به بررسی دستور وابستگی و دادگان آن پرداخته خواهد شد.

۴. دستور وابستگی

تجزیهٔ وابستگی[۵] رهیافتی برای تجزیهٔ نحوی زبان طبیعی به صورت خودکار است. این رهیافت از زبان‌شناسی سنتی مبتنی بر دستور وابستگی[۶] اقتباس شده است. در سال‌های اخیر این روش بیش از پیش مورد توجه قرار گرفته است. چند دلیل عمده برای این اقبال عمومی وجود دارد. نخست این که این گونه از نمایش ساختار نحوی زبان طبیعی، کاربردهای بسیاری در برنامه‌های مربوط به فهم زبان طبیعی از جمله ترجمهٔ خودکار[۷] و استخراج اطلاعات[۸] دارد. دومین دلیل این است که این نوع از دستور زبان (و

---

۱ *Generative Grammar*

۲ *Penn Treebank*

۳ *Constituent*

۴ *Proposition Bank*

۵ *Dependency Parsing*

۶ *Dependency Grammar*

۷ *Machine Translation*

۸ *Information Extraction*

تجزیهٔ بر مبنای آن)، در مقایسه با دستور زبان مبتنی بر عبارات[۱]، سازگاری بیش‌تری با طبیعت زبان‌های بی‌ترتیب[۲] دارد [۹]. در چنین زبانهایی قابلیت جابه‌جایی اجزای جمله وجود دارد [۱۵]؛ مثلاً «من در مدرسه کتاب را به علی دادم»؛ «من در مدرسه به علی کتاب را دادم»؛ «من به علی در مدرسه کتاب را دادم»؛ و «من کتاب را به علی در مدرسه دادم». علاوه بر این در این زبان‌ها امکان حذف اسم‌ها و ضمایر به قرینهٔ حضوری وجود دارد و یک واحد معنایی یا نحوی واحد (مانند گروه اسمی) قابلیت تکه‌تکه شدن و پخش در سطح جمله دارد [۱۵]. امّا مهم‌ترین دلیل، نتایج رضایت‌بخش حاصل از اعمال این روش در برخی از زبان‌ها با استفاده از روش‌های یادگیری خودکار[۳] بوده است [۹].

۴-۱. نظریهٔ دستور وابستگی

نظریهٔ دستور وابستگی یکی از نظریه‌های ساخت‌گرا[۴] و صورت‌گراست[۵] که اساساً در آن از طریق بررسی روابط وابستگی بین عناصر هسته و وابسته در زبان، به توصیف ساخت‌های نحوی در زبان‌های گوناگون پرداخته می‌شود [۱۶]. شاید آغاز رویکرد زبان وابستگی مربوط به اندیشه‌های زبان‌شناسی پانینی[۶] [۱۷] در مورد زبان سانسکریت باشد؛ امّا کار تنی‌یر[۷] آغازی بر استفاده از این رویکرد در زبان‌شناسی نوین است. او نخستین بار در کتاب کم‌حجمی با عنوان گفتارهایی در نحو ساختاری [18] این دیدگاه را مطرح کرد که شرح مبسوط آن پس از مرگش در کتاب مبانی نحو ساختاری [19] منتشر شد. پس از تنی‌یر، زبان‌شناسان مختلف روش‌های مختلفی را برای ارائهٔ دستور زبان وابستگی پیشنهاد داده‌اند. در همهٔ این دستورها یک فرض پایه وجود دارد. در همهٔ انواع دستور زبان وابستگی فرض بر این است که ساختار نحوی شامل واژه‌هایی است که این واژه‌ها با روابط دودویی نامتقارن با هم در ارتباط هستند. به این روابط، ارتباط وابستگی یا وابستگی گفته می‌شود [9]. 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 به کار رفته است، تعیین کرد

---

[۱] *Phrase-Based*

[۲] *Free Order*

[۳] *Machine Learning*

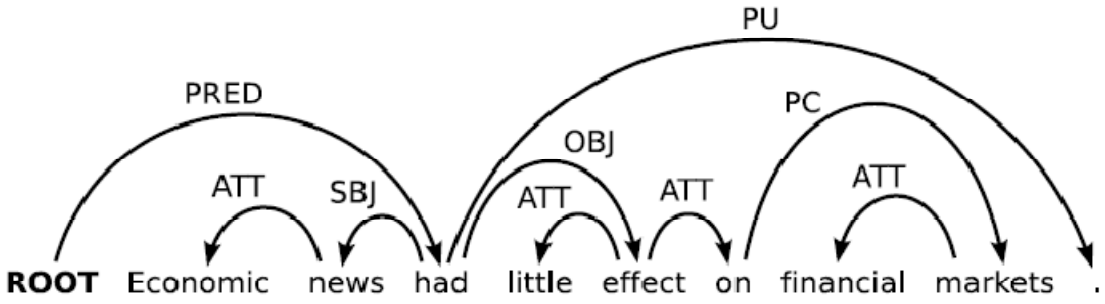
[۴] *Structuralist*

[۵] *Formalist*

[۶] *Panini*

[۷] *Tesnière*

[۱۶]. در همهٔ این رابطه‌ها یک واژه وابسته[۱] و واژهٔ دیگر سر[۲] است[۳]. در شکل ۱ نمونه‌ای از یک درخت وابستگی نشان داده شده است.



شکل ۱ نمونه‌ای از یک درخت وابستگی [9]

شایان ذکر است که اطلاعات موجود در ساختار وابستگی با اطلاعات موجود در ساختار مبتنی بر عبارات متفاوت است. در دستور وابستگی، جمله به دو بخش نهاد و گزاره تقسیم نمی‌شود. در این نظریه این که در تجزیهٔ جمله به دو گروه اسمی و فعلی تقسیم می‌شود، رد می‌شود. به اعتقاد انگل[۴] [20]، تجزیهٔ جمله به دو بخش نهاد و گزاره برای تحلیل ساخت اطلاعاتی جمله مفید است ولی در تحلیل نحوی مرکز ثقل ساختاری جمله فعل است [۱۶]. البته تبدیل اطلاعات موجود هر کدام از این دو ساختار به هم، امکان‌پذیر است ولی برای سهولت فرض بر این است که رویکرد مبتنی بر ساختار وابستگی و رویکرد مبتنی بر عبارات، دو رهیافت مختلف و متفاوت هستند [9].

۴-۲. ظرفیت

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 [۱۶]. تنی‌یر [21] مفهوم ظرفیت را از شیمی اقتباس کرده بود. ظرفیت در شیمی عبارت است از توانایی یک عنصر در ترکیب با تعداد خاصی از اتم‌های عناصر دیگر. این که ساخت بنیادین جمله حول فعل

---

[۱] *Dependent*

[۲] *Head*

[۳] اصطلاحات دیگری نیز به جای این دو اصطلاح به کار می‌رود. اصطلاح بچه (*Child*) و پیرایند (*Modifier*) به جای وابسته؛ و اصطلاح حاکم (*Governor*)، رئیس (*Regent*) و والدین (*Parent*) به جای سر به کار می‌رود.

[۴] *Engel*

مرکزی آن صورت می‌گیرد، مبین این واقعیت است که هر فعل پیش از آن که وارد جمله بشود، خود مشخص‌کنندهٔ نوع ساخت بنیادین است [۱۶].

۴-۳. فرهنگ ظرفیت واژگانی

همان‌طور که اشاره شد با استفاده از ظرفیت‌یابی می‌توان یک فرهنگ ظرفیت را ساخت که در آن ساخت ظرفیتی و متمم‌های اجباری و اختیاری اسم‌ها، فعل‌ها و صفت‌ها مشخص می‌شود. یکی از معروف‌ترین فرهنگ ظرفیت در زبان چکی است که دادگان وابستگی پراگ هم بر اساس آن ساخته شده است. یکی از این فرهنگ‌ها، فرهنگ ظرفیت فعل عربی [۲۲] از روی دادگان وابستگی عربی [۱۳] است.

۴-۴. فرهنگ‌های ظرفیت در زبان‌های دیگر[۱]

بر اساس نظر لوین [۲۳] بر اساس ساخت‌های نحوی مختلفی که در افعال وجود دارد، می‌توان ساخت معنایی‌شان را دسته‌بندی کرد. به همین صورت او، ۴۹ ردهٔ معنایی را برای افعال انگلیسی در نظر گرفت. بالدوین و همکارانش [۲۴] به طراحی واژه‌نامهٔ ظرفیتی دوسویهٔ ژاپنی-انگلیسی پرداختند که در سه سطح واژه، مفهوم[۲] و چارچوب معنایی[۳] این واژه‌نامه سطح‌بندی شده است. بنابراین در این واژه‌نامه، واژه‌ها بر اساس ریشهٔ واژه‌ای[۴]، مفهوم واژه‌ای و محتوای موضوعی[۵] خوشه‌بندی شده‌اند. این واژه‌نامه دارای ۲۹ نقش معنایی و ۷ ساختار موضوعی[۶] مختلف است. در [۲۵] علاوه بر ساختار نحوی ۳۸ هزار واژهٔ هسته در زبان انگلیسی، ساختار زیرمقوله‌های[۷] معنایی نیز درج شده است. در زبان روسی بر اساس رده‌بندی فیلمور [۲۶] در فریم‌نِت، فرهنگ واژگانی ظرفیت نحوی و معنایی ساخته شده است [۲۷]. در زبان عربی بر اساس پیکرهٔ وابستگی پراگ [۱۳]، اقدام به جمع‌آوری ظرفیت معنایی افعال شده است که در این ظرفیت ۵ نوع کنش‌گر[۸] به عنوان متمم‌ها و سه نوع نقش معنایی برای افزوده در نظر گرفته شده است [۲۲]. در [۲۸] با استفاده از روش‌های آماری در پنج پیکرهٔ معروف زبان انگلیسی، اقدام به ساخت

---

[۱] در این بخش تنها به برخی از کارهای انجام شده، اشاره شده است.

[۲] *Sense*؛ برخی اوقات این واژه به عنوان «معنا» یا «معنای واژه» (*Word Sense*) ترجمه می‌شود.

[۳] *Frame*

[۴] *Lexical Stem*

[۵] *Argument Content*

[۶] *Argument Structure*

[۷] *Subcategorization*

[۸] *Actant*

فرهنگ زیرمقوله‌های فعلی به علاوهٔ بسامد این ساخت‌ها در پیکره شده است. شبیه به کاری که در [۲۸] انجام شده، در زبان فرانسوی نیز انجام شده است [۲۹].

۵. منابع و تجزیه‌گرهای موجود

هم‌اکنون تجزیه‌گرهای مختلفی به صورت متن‌باز برای تجزیهٔ وابستگی از روی پیکره‌های متنی وجود دارد. هم‌چنین تعداد پیکره‌های وابستگی زبان‌های زندهٔ دنیا به بیش از ۱۴ زبان زنده رسیده است [۳۰] که از میان آن‌ها می‌توان پیکرهٔ درخت وابستگی عربی پراگ[۱] [۱۳] اشاره کرد. برای زبان فارسی تاکنون هیچ پیکره‌ای که در دسترس و یا خصوصی و قابل خرید باشد، وجود ندارد.

۶. مراجع

[۱] S. Abney, "Statistical methods in language processing," *Wiley Interdisciplinary Reviews: Cognitive Science,* 2010.

[۲] J. Olive*, et al.*, *Handbook of Natural Language Processing and Machine Translation: DARPA Global Autonomous Language Exploitation*: Springer, 2011.

[۳] F. Oroumchian*, et al.*, "Creating a feasible corpus for Persian POS tagging," Technical Report, no. TR3/06, University of Wollongong (Dubai Campus)2006.

[۴] R. Hwa*, et al.*, "Bootstrapping parsers via syntactic projection across parallel texts," *Natural Language Engineering,* vol. 11, pp. 311-325, 2005.

[۵] M. P. Marcus*, et al.*, "Building a large annotated corpus of English: The Penn Treebank," *Computational linguistics,* vol. 19, pp. 313-330, 1993.

[۶] P. Kingsbury and M. Palmer, "From treebank to propbank," 2002, pp. 1989–1993.

[۷] M. Palmer*, et al.*, "The Proposition Bank: An Annotated Corpus of Semantic Roles," *Association for Computational Linguistics,* vol. 31, pp. 71-105, 2005.

[۸] A. Böhmová*, et al.*, "The prague dependency treebank: Three-level annotation scenario," in *Treebanks: Building and using syntactically annotated corpora.* : Kluwer Academic Publishers, 2001.

[۹] S. Kübler*, et al.*, *Dependency Parsing*: Morgan & Claypool, 2009.

[۱۰] K. Oflazer*, et al.*, "Building a Turkish treebank ",*Treebanks: Building and Using Parsed Corpora,* vol. 20, pp. 261-277, 2003.

[۱۱] L. Van der Beek*, et al.*, "The Alpino dependency treebank," *Language and Computers,* vol. 45, pp. 8-22, 2002.

[۱] *Prague Arabic Dependency Treebank*

[۱۲] M. T. Kromann, "The Danish Dependency Treebank and the DTAG treebank tool," presented at the Second Workshop on Treebanks and Linguistic Theories (TLT 2003), Växjö, Sweden, 2003.

[۱۳] J. Hajič, *et al.*, "Prague Arabic Dependency Treebank: Development in data and tools," presented at the Proceedings of the NEMLAR 200 ۴International Conference on Arabic Language Resources and Tools, Cairo, Egypt, 2004.

[۱۴] B. McGillivray, *et al.*, "The Index Thomisticus treebank project: Annotation, parsing and valency lexicon," *Traitement Automatique des Langues,* vol. 50, 2009.

[۱۵] S. Steele, "Word order variation: a typological survey," in *Universals of human language*. vol. 4: Syntax: Stanford University Press, 1978, pp. 585-623.

[۱۶] ا. طبیب‌زاده, ظرفیت فعل و ساختهای بنیادین جمله در فارسی امروز: نشر مرکز, ۱۳۸۵.

[۱۷] A. Bharati, *et al.*, *Natural Language Processing: A Paninian Perspective*. New Delhi: Prentice Hall of India, 1994.

[۱۸] L. Tesnière, *Esquisse d'une Syntaxe structurale*. Paris: Klincksieck, 1953.

[۱۹] L. Tesnière, *Éléments de syntaxe structurale*: Editions Klincksieck, 19.۵۹

[۲۰] U. Engel, *Kurze Grammatik der deutschen Sparche*. München: Iudicium Verlge, 2002.

[۲۱] L. Tesnière, *Grundgzüge der Strukturalen Syntax*. Stuttgart: Klettcotta, 1980.

[۲۲] V. bielický and O. Smrz, "Building the Valency Lexicon of Arabic Verbs," presented at the Proceedings of the 6th Conference on Language Resources & Evaluation (LREC 2008), Marrakech, Morocco, 2008.

[۲۳] B. Levin, *English Verb Classes and Alternations*. Chicago, London: University of Chicago Press, 1993.

[۲۴] T. Baldwin, *et al.*, "A valency dictionary architecture for machine translation," in *8th International Conference on Theoretical and Methodological Issues in Machine Translation (TMI-99)*, Chester, UK, 1999, pp. 207–217.

[۲۵] R. Grishman, *et al.*, "COMLEX syntax: Building a computational lexicon," in *15th International Conference on Computational Linguistics: COLING-94*, Kyoto, Japan, 1994, pp. 268-272.

[۲۶] C. J. Fillmore, "Border Conflicts: FrameNet meets construction grammar," presented at the XIII EURALEX, 2008.

[۲۷] O. Lyashevskaya, "Bank of Russian Constructions and Valencies," presented at the Seventh conference on International Language Resources and Evaluation (LREC'10), Valletta, Malta, 2010.

[۲۸] A. Korhonen, *et al.*, "A large subcategorization lexicon for natural language processing applications," presented at the 5th LREC, 2006.

[۲۹] C. Messiant, *et al.*, "Lexschem: A large subcategorization lexicon for french verbs," presented at the LREC 2008, 2008.

[۳۰] J. Nivre, *et al.*, "The CoNLL 2007 Shared Task on Dependency Parsing," presented at the Proceeding of CoNLL 2007, New York, USA, 2007.