



پژوهشگاه علوم انسانی و مطالعات فرهنگی



انجمن زبان‌شناسی ایران

هم‌اندیشی زبان فارسی در رایاسپهر

استخراج هستان‌شناسی برای دانشنامه معماری و شهرسازی

حامد ملک

دانشگاه صنعتی امیرکبیر

یحیی تابش

دانشگاه صنعتی شریف

۱۰ دی ماه ۱۳۸۸



معرفی پروژه

- ❖ تشکیل پایگاه معرفتی عظیمی برای تاریخ معماری ایران از هزاره هفتم پیش از میلاد تا کنون
- ❖ انتشار محصول این طرح به صورت وبگاه
- ❖ دامنه جغرافیایی طرح همه جهان ایرانی (از سیحون تا خلیج فارس و از سند تا فرات) است.
- ❖ دامنه تاریخی موضوع طرح از روزگار پیش از تاریخ تا پایان دوره پهلوی اول را شامل می شود.



معرفی پروژه

- ❖ دانشنامه متعارف ابزاری است که پیشینه تحقیق درباره عنوانی معین به نام «مدخل» را در اختیار عموم یا پژوهشگران قرار می دهد.
- ❖ در این دانشنامه همه محققان مؤلف بالقوه مقاله درباره هر مدخل شمرده می شوند.
- ❖ در این دانشنامه مداخل ثابت نیست و به ازای جستجو و نیاز مخاطبان، پیوسته بر مداخل افزوده می شود.
- ❖ اساس کار در این دانشنامه امکان جستجوی دقیق در همه مدارک مرتبط با تاریخ معماری و شهرسازی ایران زمین است.
- ❖ مجموعه اطلاعات به دو دسته «مدارک» و «مداخل» تقسیم می شوند.
- ❖ مدخل: واژه ای که اطلاعات حول آن گردآوری یا تولید می شود.
- مدرک: هر نوع منبع چاپی یا غیر چاپی مانند کتاب، فیلم، نسخه خطی و غیره است.



معرفی پروژه

- ❖ مجری این طرح فرهنگستان هنر و حامی اصلی آن وزارت مسکن و شهرسازی است.
- ❖ همکاران طرح: سازمان میراث فرهنگی، سازمان اسناد و کتابخانه ملی و دانشگاه شهید بهشتی
- ❖ طرح اکنون در مرحله تأسیس قرار دارد که هدف آن تهیه ۵۰۰۰ مدخل اولیه در وبگاه است.
- ❖ در مرحله کنونی زبان دانشنامه فارسی و انگلیسی است که در آینده به همه زبان هایی که اطلاعاتی در آن زبان وجود دارد گسترش خواهد یافت.



معرفی پروژه

<http://www.eiah.org>



دانشنامهٔ تاریخ معماری و شهرسازی ایران زمین
ENCYCLOPEDIA OF IRANIAN ARCHITECTURAL HISTORY

اصطلاح
کاری

آثار
مسجد کوفیه

نام جغرافیایی
سند

دوره تاریخی
مخاندین

منبع اولیه
الزبیلاد و احبارالعقاد

طرح دانشنامه تاریخ معماری و شهرسازی ایران زمین

دانشنامهٔ تاریخ معماری و شهرسازی ایران زمین بر مبنای تعریفی تازه از «دانشنامه» شکل گرفته است. دانشنامه‌ها در معرفت‌های گوناگون بشری محصول تراکم نتایج فعالیت‌های پژوهشی است و دانشنامهٔ متعارف ابزاری است که مجموعهٔ پیشینهٔ تحقیقها دربارهٔ عنوانی معین به نام «مدخل» را در اختیار عموم یا پژوهشگران قرار می‌دهد و امکان آشنایی عموم و ادامهٔ کار پژوهشگران را فراهم می‌آورد. در دانشنامه‌های متعارف، مجموعهٔ سابقهٔ تحقیق دربارهٔ مدخل در اختیار محققین برجسته قرار می‌گیرد و او مقاله‌های دربارهٔ آن مدخل تألیف می‌کند.

ادامه ...

پایگاه دانشنامه در دست تأسیس است و آنچه در پیش رو دارید صرفاً نسخهٔ آزمایشی وبگاه برای عرضهٔ تصویری از آیندهٔ آن است. دانشنامه از همهٔ دستگامها و اشخاص حقیقی و حقوقی علاقه‌مند و متولیان ارشیهای خصوصی و عمومی دعوت می‌کند در این کار بزرگ ملی و منطقه‌ای مشارکت کنند.

تازه‌ها

طرح معرفی مدخل‌ها

از مهم‌ترین طرح‌های در دست انجام واحد پژوهش دانشنامه، گزینش و تعیین مداخل در هر هفت حوزهٔ آثار، اعلام جغرافیایی، اصطلاحات، اشخاص، دوره‌های تاریخی، منابع اولیه، و آثار منقول است که در درازمدت انجام خواهد شد. پیش از انعام کامل طرح مدخل‌گزینی، دانشنامه در نظر دارد در مقاطع زمانی کوتاه‌تر تعدادی از مداخل هر یک از این حوزه‌ها را از طریق وبگاه در دسترس کاربران قرار دهد، بدین ترتیب حدود ۲۵۰ مدخل گزینش‌شده در حوزهٔ آثار، مشتمل بر **ادامه ...**

طرح معرفی کتاب

تا پایان اسفند امسال، ۲۰۰ عنوان کتاب فارسی در حوزهٔ تاریخ معماری و شهرسازی ایران زمین در وبگاه دانشنامه معرفی خواهد شد. این کار با هدف ارائهٔ انواع مدارک نوشتاری صورت می‌گیرد. کتاب‌های مورد نظر شامل منابع تاریخی، جغرافیای تاریخی، سفرنامه‌ها، و سایر کتب معتبر معاصر است. اطلاعاتی که برای معرفی هر کتاب ارائه می‌شود، شامل **ادامه ...**

کتاب حقوق برای نویسندگان نشر جمهوری اسلامی محفوظ است.

صفحه اصلی

جستجو

جستجوی پیشرفته

مخزن اطلاعات دانشنامه

- درباره دانشنامه
- اختیار دانشنامه
- راههای کاربران
- تماس با دانشنامه
- انجمن

مدخل‌ها

- اصطلاح
- آثار
- نام جغرافیایی
- اشخاص
- دوره تاریخی
- منبع اولیه
- آثار منقول

مدرک‌ها

- نوشته
- عکس
- طرح
- مدرک شنیداری
- چند رسانه‌ای

همکاران




English

فارسی

سازمان اسناد و کتابخانه ملی جمهوری اسلامی ایران

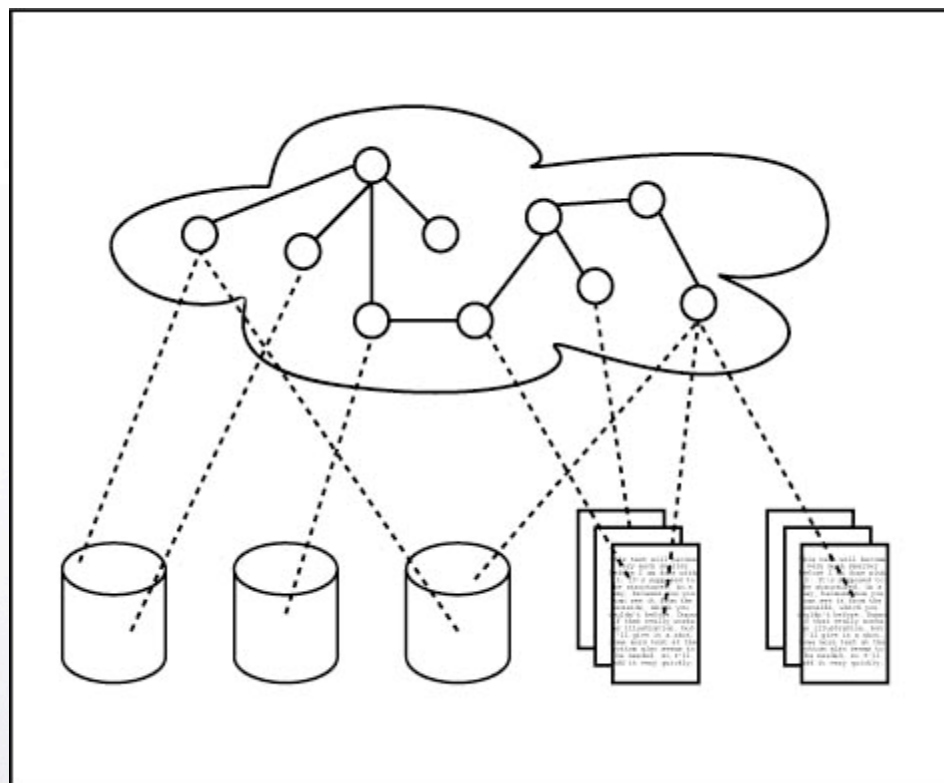


معماری اطلاعاتی

- ❖ استاندارد مورد استفاده برای بازنمایی و تبادل دانش استاندارد «مدل موضوعی» یا Topic Maps است.
- ❖ در این مدل، اطلاعات با استفاده از سه مفهوم موضوعات، وابستگی ها و رخدادها بازنمایی می شوند.
- ❖ موضوع: نماینده یک مفهوم مانند شخص، مکان، رویداد و غیره
وابستگی: نحوه ارتباط بین موضوعات
رخداد: نمایش رابطه بین موضوعات و منابع اطلاعاتی مرتبط
- ❖ مدل موضوعی را می توان نوع خاصی از فناوری وب معنایی دانست.

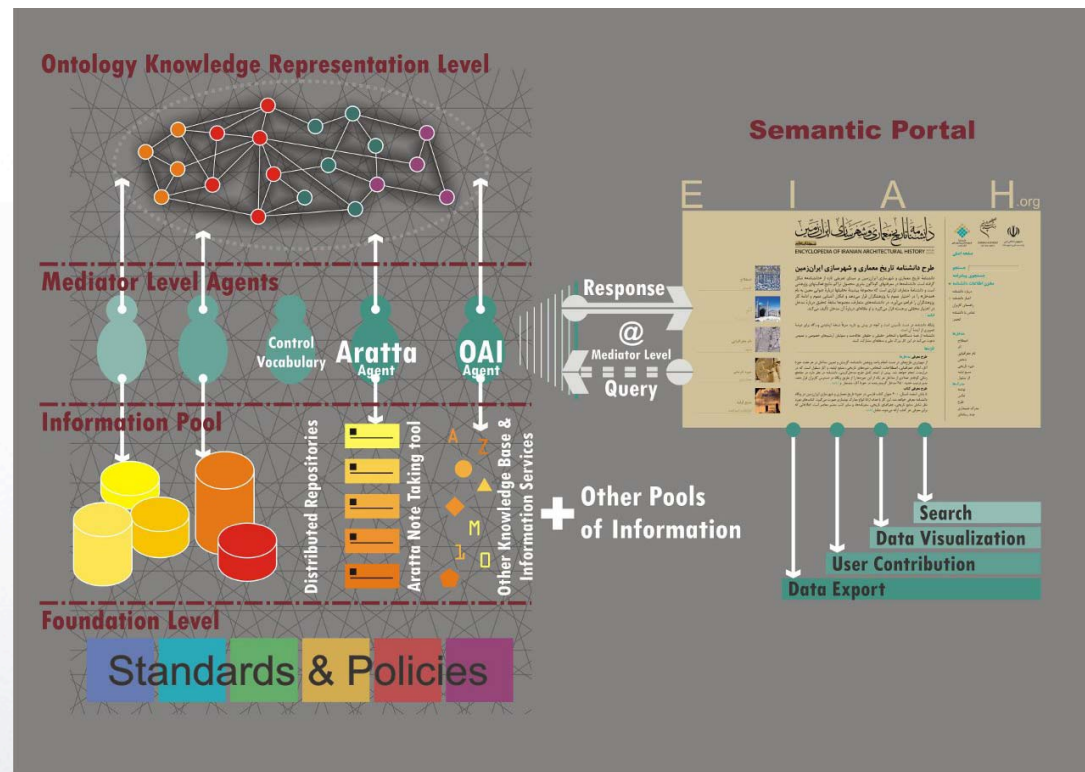


معماری اطلاعاتی





معماری اطلاعاتی





هستان شناسی

هستان شناسی سنگین

هستان شناسی سبک

مدل های مفهومی

(تعریف طبقات، ویژگی و روابط آنها)



معماری اطلاعاتی

پروژه آرته

جستجوی تقریبی

پروژه استخراج اسامی خاص

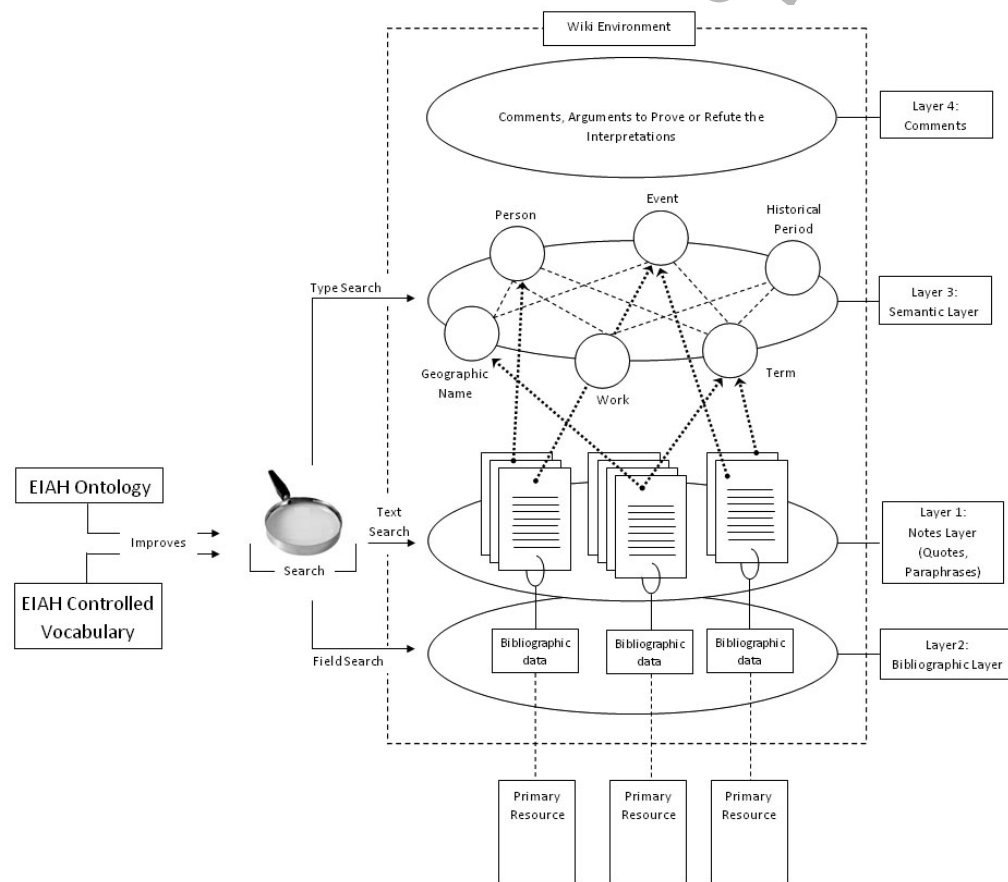


پروژه آرته

- ❖ هدف: ایجاد یک کتابخانه معنایی از یادداشت های تحقیقاتی
- ❖ نام پروژه برگرفته از نام تمدن باستانی گمشده ای به نام آرته است.
- ❖ بیش از ۱۰ هزار یادداشت تحقیقاتی



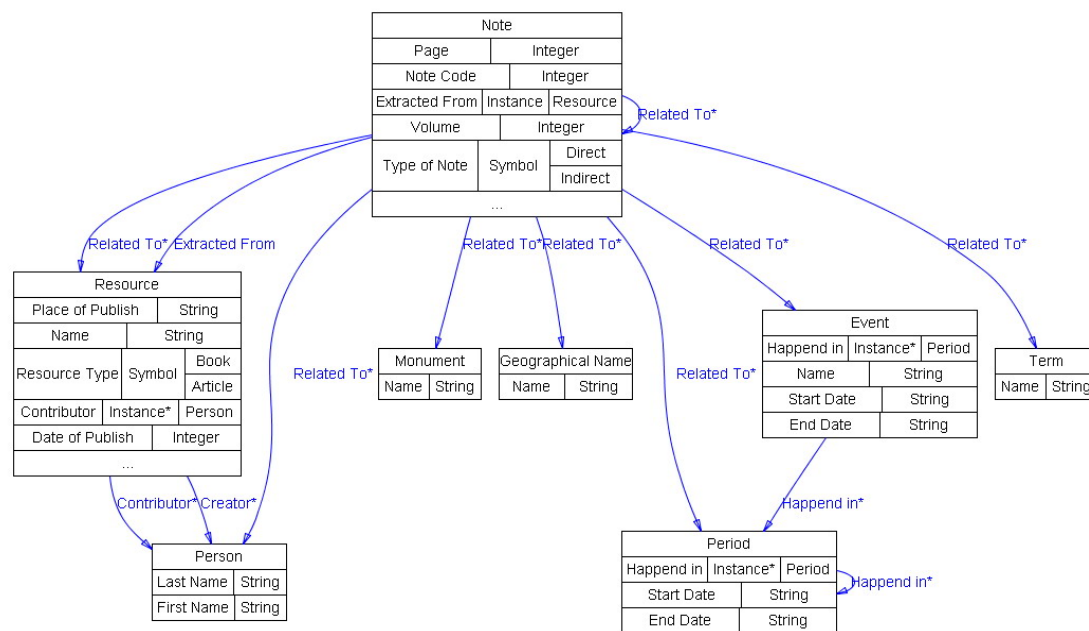
پروژه آرته



نمودار مدل داده



پروژه آرته



هستان شناسی آرته



پروژہ آرٹہ

❖ مڈیاویکی (Mediawiki)

❖ سمنٹیک مڈیاویکی (Semantic Mediawiki)

❖ سمنٹیک فرم (Semantic Form)

❖ استاندارد ڈابلین کور (Dublin Core)



جستجوی تقریبی

- ❖ هدف: یافتن عبارات و زیررشته هایی که تفاوت قابل اغماض با رشته مورد جستجو دارند.
- ❖ جستجوی تقریبی بی نقص نیاز وسیع به اطلاعات زبانی، واژگانی، معنایی و مانند آن دارد.
- ❖ در این پروژه به پیاده سازی شیوه هایی که توسط الگوریتم ها و جداول ساده قابل پیاده سازی هستند پرداخته شده است.
- ❖ برای این منظور از توصیف «جستجوی تقریبی برای زبان فارسی ایران» تهیه شده توسط شرکت فارسی وب شریف استفاده شده است.



جستجوی تقریبی

- ❖ غلط های کدگذاری: مانند استفاده از نااستاندارد از حروف عربی به جای حروف فارسی
- ❖ غلط های املائی معمول یا نامحسوس: مانند نوشتن کلمه «مؤمن» به شکل «مومن»
- ❖ تفاوت در اعراب گذاری عبارات و وجود یا عدم وجود نویسه های ضمنی
- ❖ همه موارد فوق در قالب این پروژه به موتور جستجوی آزاد لوسن «Lucene» پیاده سازی شده و قابل استفاده برای عموم است.



تشخیص اسامی خاص

- ❖ با توجه به حجم بالای متون نوشتاری و نیاز به برچسب گذاری آنها برای تولید مناسب هستان شناسی، استفاده از روش های خودکار یک الزامی است.
- ❖ ترجیح روش های آماری بر روش های زبان شناسی به دلیل استقلال روش از زبان
- ❖ انتخاب کتابخانه معنای آرته به عنوان منبع اصلی استخراج اسامی خاص
- ❖ پیاده سازی الگوریتم با کمک ابزارهای آزاد موجود



تشخیص اسامی خاص

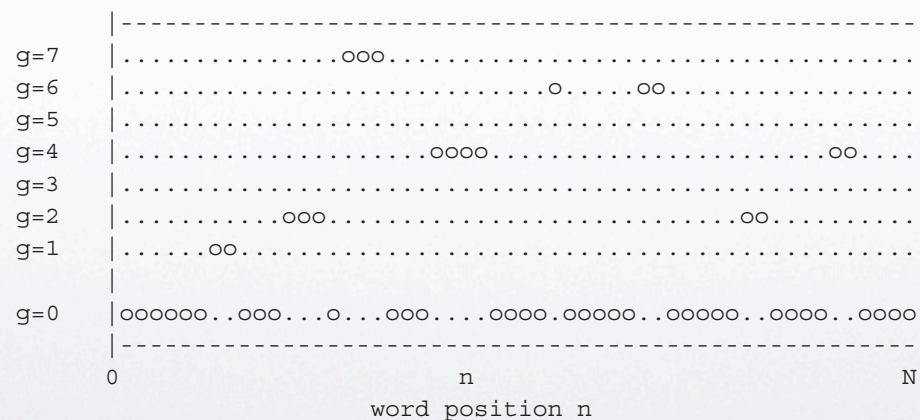
- ❖ اسامی به ۶ دسته یا کلاس تقسیم می شوند.
- ❖ اصطلاح: آتشگاه، خاتم کاری، کاروان سرا
- ❖ اثر: معبد آناهیتا، مسجد جامع اصفهان، حمام علی قلی آقا
- ❖ نام جغرافیایی: آباد، دریای مازندران، کرمان
- ❖ دوره تاریخی: آل بویه، ساسانیان، صفویه
- ❖ شخص: سلطان سنجر، لطف علی خان زند، شاه عباس
- ❖ منبع اولیه: آثار عجم، سلجوق نامه، سفرنامه ناصر خسرو



تشخیص اسامی خاص

$w_1^N = w_1 \dots w_n \dots w_N$ دنباله ورودی:

$g_1^N = g_1 \dots g_n \dots g_N$ دنباله خروجی:





تشخیص اسامی خاص

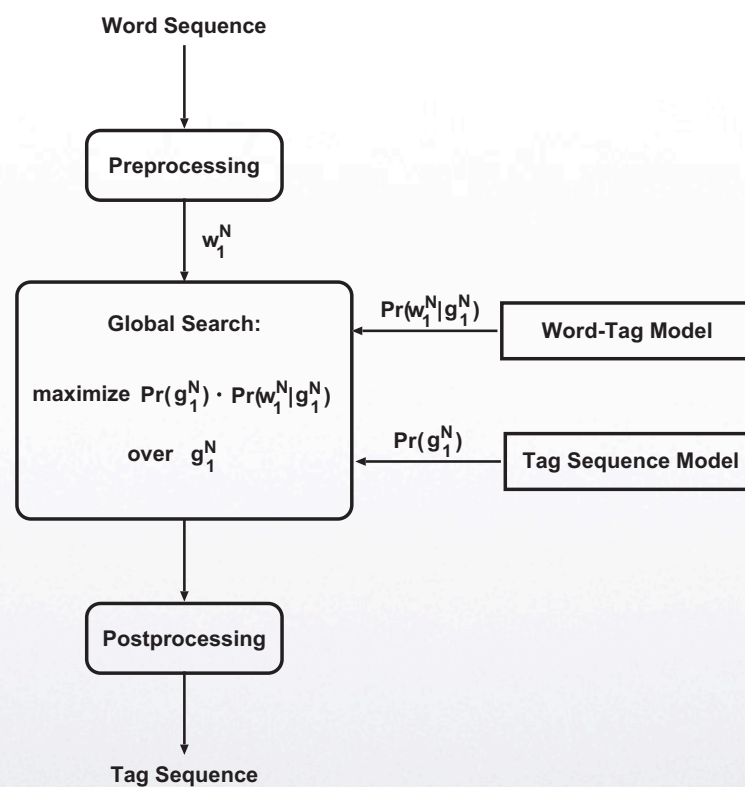
ایده اصلی مشابه روش معمول آماری برای POS Tagging است.

قانون تصمیم گیری بیز:

$$\begin{aligned}w_1^N \rightarrow \hat{g}_1^N &= \arg \max_{g_1^N} Pr(g_1^N | w_1^N) \\ &= \arg \max_{g_1^N} \{Pr(g_1^N, w_1^N)\} \\ &= \arg \max_{g_1^N} \{Pr(g_1^N) \cdot Pr(w_1^N | g_1^N)\}\end{aligned}$$



تشخیص اسامی خاص





تشخیص اسامی خاص

$$\arg \max_{g_1^N} Pr(g_1^N | w_1^N) = \arg \max_{g_1^N} \{Pr(g_1^N) \cdot Pr(w_1^N | g_1^N)\}$$

احتمال تعلق: مدل مرتبه صفر

$$Pr(w_1^N | g_1^N) = \prod_{n=1}^N p_0(w_n | g_n)$$

احتمال بایگرام POS: مدل مرتبه یک

$$Pr(g_1^N) = \prod_{n=1}^N p_1(g_n | g_{n-1})$$

آموزش: هر دو توزیع احتمالاتی بالا را با کمک نمونه های دستی تخمین می زنیم.



تشخیص اسامی خاص

$$\begin{aligned} p_0(w|g) &= \frac{p(w, g)}{p(g)} \\ &= \frac{p(w) \cdot p_0(g|w)}{p(g)} \end{aligned}$$

مسأله بهینه سازی نهایی:

$$\arg \max_{g_1^N} Pr(g_1^N | w_1^N) = \arg \max_{g_1^N} \left\{ \prod_{n=1}^N \left[\frac{p_0(g_n | w_n)}{p(g_n)} \cdot p_1(g_n | g_{n-1}) \right] \right\}$$

برای حل مسأله بهینه سازی بدست آمده از روش برنامه ریزی پویا استفاده می کنیم.



تشخیص اسامی خاص

- ❖ تفاوت اصلی با POS Tagging در طول برچسب ها است که می تواند بیش از یک باشد.
- ❖ برای این منظور از یک مدل ساده طول استفاده می شود:
- ❖ هر برچسب با دو برچسب «شروع» یا «ادامه» جایگزین می شود.



تشخیص اسامی خاص

قانون تصمیم گیری بیز:

$$\begin{aligned}w_1^N \rightarrow \hat{g}_1^N &= \arg \max_{g_1^N} Pr(g_1^N | w_1^N) \\ &= \arg \max_{g_1^N} \{Pr(w_1^N) \cdot Pr(g_1^N | w_1^N)\} \\ &= \arg \max_{g_1^N} \{Pr(g_1^N, w_1^N)\}\end{aligned}$$

$$\begin{aligned}Pr(g_1^N, w_1^N) &= \prod_n Pr(g_n, w_n | g_1^{n-1}, w_1^{n-1}) \\ &= \prod_n [Pr(g_n | g_1^{n-1}, w_1^{n-1}) \cdot Pr(w_n | g_1^n, w_1^{n-1})] \\ Pr(g_1^N, w_1^N) &= \prod_n Pr(g_n, w_n | g_1^{n-1}, w_1^{n-1}) \\ &= \prod_n [Pr(g_n | g_1^{n-1}, w_1^{n-1}) \cdot Pr(w_n | g_1^n, w_1^{n-1})]\end{aligned}$$



تشخیص اسامی خاص

قانون تصمیم گیری بیز نهایی:

$$\arg \max_{g_1^N} Pr(g_1^N | w_1^N) = \arg \max_{g_1^N} \left\{ \prod_n [p(g_n | g_{n-1}, w_{n-1}) \cdot p(w_n | g_{n-1}^n, w_{n-1})] \right\}$$

فرض اضافه:

$$p(w_n | g_{n-1}^n, w_{n-1}) = \begin{cases} p(w_n | g_n) & g_n = g_{n-1} \\ p(w_n | g_n, w_{n-1}) & g_n \neq g_{n-1} \end{cases}$$



تشخیص اسامی خاص

500	تعداد فیش های آموزش
56343	تعداد کلمات آموزش
100	تعداد فیش های آزمایش
12330	تعداد کلمات آزمایش

2862	اصطلاح
1347	اثر
842	نام جغرافیایی
107	دوره تاریخی
392	شخص
101	منبع اولیه



کارهای آینده

- ❖ هوشمندسازی الگوریتم های جستجو با کمک روش های پردازش زبان طبیعی
- ❖ استفاده از تکنیک های آماری بهبود کیفیت تشخیص اسامی خاص
- ❖ استفاده از تکنیک های زبان شناسی در الگوریتم تشخیص اسامی خاص



اطلاعات بیشتر

❖ وبگاه دانشنامه معماری و شهرسازی ایران زمین: www.eiah.org

❖ وبگاه پروژه آرته: <http://sohrab.eiah.org:3080/aratta>

hmalek@aut.ac.ir

tabesh@sharif.ir